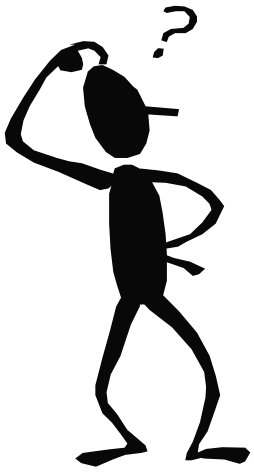
Traitement statistique des données zootechniques et sanitaires

Les méthodes d'analyses factorielles et de classification

Samir MESSAD



« Les problèmes ne viennent pas tant de ce que l'on ignore, mais de ce que l'on sait. »

Artemus Ward.

Avant-propos

Ce modeste aperçu se veut être avant tout, une manière de présenter l'analyse exploratoire et descriptive des données. Reprenant la première partie du document de B. Faye et largement inspiré de quelques ouvrages ou supports de cours en la matière, il s'agira d'exposer à des débutants ou à des personnes ayant eu une première expérience de ces méthodes, les notions de base en jeu et d'expliquer le plus simplement possible, leur utilité et mise en œuvre. Nous nous affranchirons des formules mathématiques sauf quelques exceptions indispensables, pour un exposé plus qualitatif. Ce qui ne nous empêchera pas de détailler les mécanismes de calcul indispensables à comprendre afin de pouvoir interpréter soi-même une analyse. Au final, nous espérons que le lecteur trouvera au travers des méthodes qui sont exposés dans ce document et du cours qui l'accompagne, les rudiments d'un apprentissage plus général du traitement statistique.

Les analyses ont été réalisées avec les fonctions de statistiques multivariées des librairies mva (multivariate analysis, R Core Team) et ade4 (Thioulouse et al., 2002) sous l'environnement statistique et graphique **R**. Le lecteur trouvera dans la dernière partie de ce document les informations utiles pour se procurer ces outils et la documentation.

Table des matières

1. INTRODUCTION 6

2. UN EXEMPLE POUR COMPRENDRE LES NOTIONS EN JEU DANS LES METHODES FACTORIELLES 7

2.1. LA NOTION DE STRUCTURE DANS UN TABLEAU DE DONNEES 7
2.2. EXEMPLE A 2 DIMENSIONS 7
2.3. EXEMPLE A 3 DIMENSIONS 10
2.4. LA NOTION DE DISTANCE 12

3. LES METHODES FACTORIELLES D'ANALYSE DES DONNEES 14

3.1. DIFFERENTES METHODES POUR DIFFERENTS TABLEAUX DE DONNEES 14
3.2. L'ANALYSE EN COMPOSANTES PRINCIPALES (ACP) 15
3.2.1. Objectifs 15
3.2.2. Une transformation du tableau de données 16
3.2.3. Analyse du nuage des individus 17
3.2.4. Analyse du nuage des variables 19
3.2.5. La dualité 20
3.2.6. Aides à l'interprétation 20
3.2.7. Exemple : caractéristiques pondérales des carcasses de bovins 24
3.2.8. En résumé 29

4. L'ANALYSE FACTORIELLE DES CORRESPONDANCES 29

4.1. LE TABLEAU DE CONTINGENCE 29
4.2. TRANSFORMATION DU TABLEAU DE DONNEES 30
4.3. LA RESSEMBLANCE ENTRE PROFILS 31
4.4. CONSTRUCTION DES NUAGES ET AJUSTEMENT 32
4.5. LA DUALITE 33
4.6. INTERPRETATION D'UNE AFC 34
4.7. L'INERTIE 35

5. UNE METHODE POUR TRAITER LES DONNEES D'ENQUETE : L'ANALYSE FACTORIELLE DES CORRESPONDANCES MULTIPLES (AFCM) 36

5.1. LES DONNEES 38
5.2. APUREMENT ET HOMOGENEISATION 38
5.3. TRANSFORMATION DU TABLEAU DE DONNEES 39
5.4. CONSTRUCTION DES NUAGES ET AJUSTEMENT 40
5.5. L'INTERPRETATION DES RESULTATS 41
5.5.1. Le diagramme des valeurs propres 42
5.5.2. Les aides numériques à l'interprétation 43
5.5.3. Les représentations graphiques 44
5.5.4. Description des facteurs 49
5.5.5. Les éléments supplémentaires 50

6. LES METHODES AUTOMATIQUES DE CLASSIFICATION 53

6.1. PRINCIPE 53
6.2. LA CLASSIFICATION ASCENDANTE HIERARCHIQUE (CAH) 54
6.2.1. Quel tableau soumettre à l'analyse ? 54
6.2.2. Le principe de l'algorithme de regroupement 55
6.2.3. Choix du critère d'agrégation 55
6.2.4. CAH à partir des coordonnées factorielles 57
6.2.5. Le diagramme des indices de niveau 59
6.2.6. L'arbre hiérarchique ou dendrogramme 60
6.2.7. La description des classes 60
6.2.8. La complémentarité analyse factorielle et classification 63

7. POUR ALLER PLUS LOIN EN ANALYSE DES DONNEES 65

8. BIBLIOGRAPHIE 67

8.1. LIENS INTERNET 68
8.2. L'AIDE EN LIGNE DES FONCTIONS DES LIBRAIRIES ADE4 ET MVA. 68
8.3. LISTE DES FONCTIONS DE LA LIBRAIRIE ADE4 69

Table des illustrations

Figure 1. Nuage de points. 8
Figure 2. Positions des individus sur l'axe factoriel. 8
Figure 3. Ajustement des deux dimensions du nuage de points. 9
Figure 4. Positions des individus A, B et E dans le nuage à 3 dimensions. 11
Figure 5. Exploration de la structure du tableau 2 par une série de plans. 11
Figure 6. Projection orthogonale sur F1 de la distance entre les points C et I. 13
Figure 7. Déformation de la distance entre deux points par la projection orthogonale. 13
Figure 8. Illustration de la réduction des variables dans le calcul de la distance entre A et B. 17
Figure 9. Recherche des directions principales du nuage par les axes factoriels. 18
Figure 10. Projections des points sur l'espace factoriel F1-F2. 18
Figure 11. Représentation graphique des variables normées. 19
Figure 12. Projection orthogonale du point-variable X^j sur le facteur. 20
Figure 13. Diagramme des valeurs propres. 21
Figure 14. Contribution d'un point à l'inertie du nuage. 21
Figure 15. L'effet taille en ACP. 22
Figure 16. Mesure de la qualité de la représentation des points M et N par le cos² de l'angle α. 23
Figure 17. Individu très contributif mis en supplémentaire. 24
Figure 18. Diagramme des valeurs propres de l'ACP normée du tableau zebus. 25
Figure 19. Représentation des bovins et des caractéristiques pondérales sur le plan factoriel 1-2. 26
Figure 20. Représentation des bovins et des caractéristiques pondérales sur le plan 1-3. 27
Figure 21. Diagrammes de dispersion facteur 1 (en abscisse) et variables initiales (en ordonnée). 28
Figure 22. Représentation du lien entre race et les facteurs 1 et 2 de l'ACP de Zebus. 28
Figure 23. Transformations du tableau de contingence K. 31
Figure 24. Diagramme des valeurs propres de l'AFC du tableau regicu. 33
Figure 25. Représentation simultanée sur plan factoriel 1-2 des régions et des cultures. 34
Figure 26. Représentation des cultures et des régions sur le premier facteur. 34
Figure 27. Diagramme des valeurs propres. 35
Figure 28. Diagramme des valeurs propres de l'ACM du tableau eleveurs. 42
Figure 29. L'effet Guttman en AFC ou AFCM. 42
Figure 30. Différentes configurations de la répartition des individus sur les plans factoriels. 45
Figure 31. Plan factoriel 1-2 des individus 45
Figure 32. Le plan factoriel 1-2 des modalités. 46
Figure 33. Synthèse des scores des modalités sur le facteur 3 à l'aide de la fonction graphique score.acm. 47
Figure 34. Synthèse des scores des modalités sur le plan factoriel 1-3 à l'aide de la fonction graphique scatter.acm. 48
Figure 35. Deux variables qualitatives associées sur le plan factoriel 1-2. 48
Figure 36. Plan factoriel 1-2 des modalités actives et supplémentaires (taille des caractères plus grande). 51
Figure 37. Représentation des liens entre variables supplémentaires et les facteurs 1 et 2 de l'ACM des éleveurs 52
Figure 38. Représentation d'une hiérarchie de partitions. 53
Figure 39. Partition finale obtenue par méthode nuées dynamiques. 53
Figure 40. Agrégation d'un individu à la classe G_1 et diminution de l'inertie inter. 57
Figure 41. Synoptique d'une analyse typologique. 58
Figure 42. Diagramme des indices de niveau d'une hiérarchie. 59
Figure 43. L'arbre hiérarchique ou dendrogramme. 60
Figure 44. Représentation des classes de la partition des éleveurs sur le plan 1-2 des individus. 61
Figure 45. Description à l'aide des valeurs-tests de la la partition des éleveurs à l'aide des des réponses aux questions sur les problèmes de développement (variables supplémentaires). 63

Tableau 1. Exemple de tableau à 2 variables. 7
Tableau 2. Exemple de tableau à 3 variables. 10
Tableau 3. Exemple de tableau de contingence. 14
Tableau 4. Dictionnaire des variables de l'enquête éleveurs du Breedland. 37

Équation 1. Notations des distributions de fréquences. 31
Équation 2. Distances entre 2 profils lignes. 32
Équation 3. Distance entre 2 profils colonnes. 32
Équation 4. Distance entre 2 points individus issus d'un tableau disjonctif complet. 41
Équation 5. Distance entre 2 points modalités issus d'un tableau disjonctif complet. 41
Équation 6. Décomposition de l'inertie (relation de Huygens). 56
Équation 7. Formules des composantes de l'inertie. 57

Image 1. L'analyse factorielle propose divers points de vue sur les données. 12
Image 2. Les provinces du Breedland. 36

1. Introduction

Depuis une vingtaine d'années se développe une branche de la statistique que l'on appelle **l'analyse des données**. Celle-ci regroupe un ensemble de méthodes permettant d'explorer de vastes ensembles de données, aussi les nomme-t-on également méthodes multidimensionnelles car elles ne se contentent pas d'appréhender des phénomènes à l'aide d'une ou plusieurs variables prises séparément mais au travers d'un nombre important de variables prises simultanément.

Même si les principes fondamentaux de ces méthodes ont été découverts depuis longtemps (Pearson 1901, Hotelling 1933), c'est leur coût en moyens de calculs qui autrefois freinait les utilisateurs. Aujourd'hui grâce aux outils de calculs informatiques toujours plus puissants et à une « vulgarisation » de ces méthodes, l'éventail des utilisateurs est de plus en plus large et ce, dans les domaines scientifiques les plus variés.

Parmi ces méthodes, **les analyses factorielles** sont les plus utilisées. Leurs objectifs tiennent en un mot : synthétiser, c'est à dire explorer un tableau où l'on a consigné des individus décrits par diverses variables (descripteurs) de même nature. Cette exploration consiste à décrire la structure d'organisation des données en mettant en avant les regroupements, les oppositions, les tendances, les interrelations qui peuvent exister dans un ensemble d'observations.

Les méthodes de classification automatique constituent le deuxième groupe de méthodes de l'analyse des données. Elles mettent en œuvre des algorithmes permettant de grouper des ensembles d'individus ou observations en catégories homogènes au regard d'une série de descripteurs. Elles se placent en outil complémentaire de l'analyse factorielle. En effet après avoir étudié les grands traits distinctifs des individus, on voudrait pouvoir les regrouper au vu de leurs similitudes.

Ces méthodes ont un atout considérable. Il tient au fait qu'un minimum de formalisme mathématique est nécessaire pour interpréter une analyse (en tous cas en ce qui concerne les analyses de base). En effet, les notions géométriques de distance entre objets fondamentales à l'analyse factorielle s'appréhendent assez intuitivement. De plus, l'utilisation de ces méthodes se fait sans référence à des hypothèses de nature statistique ou à un modèle particulier (loi de probabilité), ce qui les distinguent de la statistique inférentielle.

On voudrait préciser dans cette introduction que le pouvoir explicatif de ces méthodes est limité et que leur vocation première est descriptive, description qui s'appuie sur une connaissance et des hypothèses préalables faites sur les données. Enfin qu'il n'y a pas de méthode exacte de dépouillement d'enquête mais une démarche (que l'on va tenter d'exposer), fondée sur l'expérience des praticiens de l'analyse des données.

2. Un exemple pour comprendre les notions en jeu dans les méthodes factorielles

2.1. La notion de structure dans un tableau de données

Avant de parler de structure dans un tableau de données, fixons-nous les idées au sujet de la structure d'un tableau de données. En effet, nous emploierons tout au long de ce document le terme de tableau de données qui désigne le « matériel » dans lequel sont stockées les informations à analyser. Un tableau de données comprend des lignes, usuellement indexées *i*, et des colonnes *j*. Les lignes représentent des individus statistiques ou observations (animal, exploitation agricole, produit, ...) et les colonnes représentent les variables, c'est à dire les mesures ou caractères qui décrivent l'individu (l'âge de l'animal, la surface de l'exploitation, la couleur du produit, ...).

Nous avons abordé en introduction de ce document la notion d'information principale, distinctive des individus. C'est précisément ce que l'on désignera par structure qui décrite pourra nous permettre de comprendre comment s'organise les différences majeures entre individus statistiques. Le reste de l'information est de second ordre, marginal par rapport à la structure car n'apportant pas grand chose dans la différenciation des individus. Il peut dans certains cas constituer ce que l'on a coutume d'appeler les fluctuations d'échantillonnage. Nous allons essayer d'appréhender cette notion de structure plus en détail et voir comment on peut la mettre en évidence au travers d'un exemple simple.

2.2. Exemple à 2 dimensions

Le tableau de données Poids-Taille, que l'on peut représenter graphiquement dans un espace à deux dimensions (le poids en abscisse et la taille en ordonnée).

Tableau 1. Exemple de tableau à 2 variables.

Individu	Poids	Taille
A	45	151
B	95	197
C	56	168
D	73	178
E	69	171
F	84	182
G	56	165
H	82	172
I	67	183
J	77	162

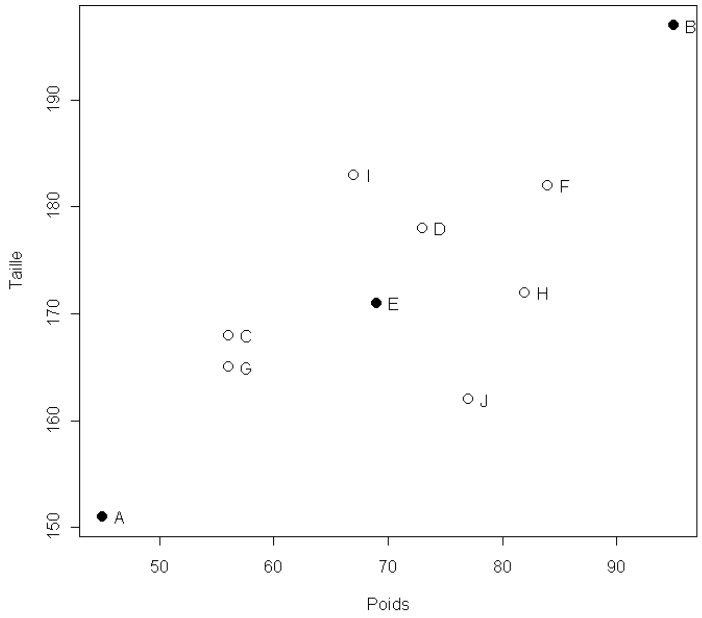


Figure 1. Nuage de points.

Cette représentation simultanée des deux dimensions (Figure 1) de notre tableau nous permet de voir comment nos données s'organisent. Ainsi, les individus petits et maigres se situent en bas et à gauche. A l'inverse, les individus grands et gros se situent en haut et à droite du nuage. Si l'on place tous les individus sur une ligne qui **résume** au mieux l'information de ce nuage, les individus A et B se situeraient aux extrémités de cette droite.

Au lieu de conserver le nuage de points, on conserverait simplement la droite qui passe au milieu du nuage, dans sa plus grande largeur, de façon à minimiser l'écart de chacun des points à cette droite. Avec deux variables, c'est une pratique courante, on appelle cette ligne, la droite de régression, elle permet d'évaluer le lien entre les variables poids et taille. Cette droite nous permet de représenter la répartition des 10 individus sur un axe synthétique (Figure 2). Synthétique, puisque la position d'un individu sur cet axe est une **combinaison** entre sa taille et son poids.

Comment interpréter les positions relatives des individus sur cet axe ?

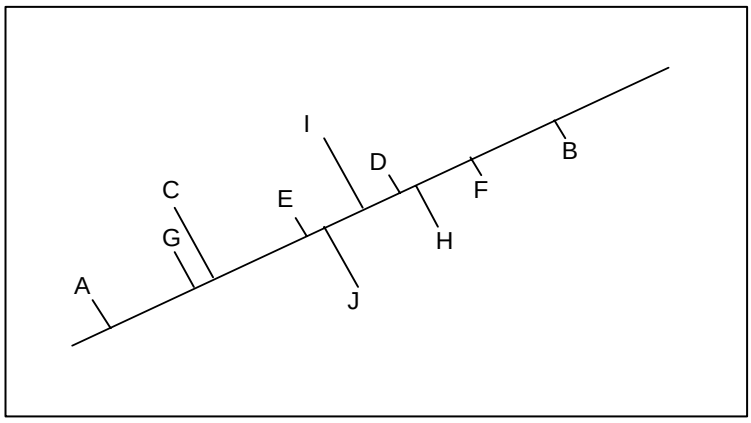


Figure 2. Positions des individus sur l'axe factoriel.

Comme il oppose les « petits-maigres » aux « grands-gros », on peut lui donner plusieurs interprétations selon les connaissances que l'on a des données d'origine. Imaginons que nous avons à faire à deux populations, l'une pygmée, l'autre nord-américaine, on pourrait interpréter l'opposition observée comme étant liée à un facteur géographique ou nutritionnel. On parlera désormais d'axe factoriel pour désigner cette droite « résumé ».

En même temps que l'on réduit le nuage en une droite, on perd irrémédiablement de l'information. En effet, on a remplacé la valeur de l'observation par une nouvelle valeur ajustée qui tente de l'approcher au mieux. La droite ajustant le nuage est la droite **d'allongement maximum**, elle tente ainsi de reproduire en simplifiant, une dimension de ce nuage de points. Si l'on veut reconstituer l'ensemble de l'information, il faudrait trouver un second axe qui passe au plus près des points en vérifiant deux contraintes :

1. Ce deuxième axe passe par l'individu moyen pour toutes les tailles et tous les poids, c'est à dire le **centre de gravité** du nuage que l'on appelle communément G (à ne pas confondre avec l'individu G de notre jeu de données).

2. Il est perpendiculaire (on dit aussi orthogonal) au premier axe factoriel.

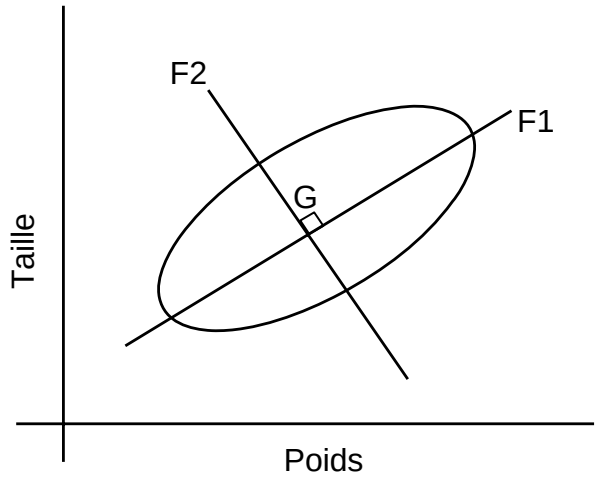


Figure 3. Ajustement des deux dimensions du nuage de points.

On ajuste ainsi la seconde dimension du nuage (Figure 3),

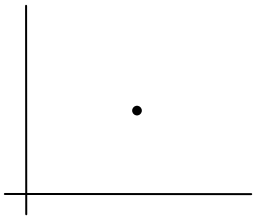
Les points possèdent désormais de nouvelles coordonnées dans le nouveau système d'axes (F1, F2) qui relate la dispersion des individus autour de l'individu moyen en taille et en poids. On appelle ces axes, les axes principaux d'inertie, l'inertie mesurant la dispersion des points par rapport au centre de gravité (nous reviendrons sur cette notion un peu plus loin).

Dégager la structure d'un tableau de données, c'est finalement expliquer les traits caractéristiques du nuage de points associé, on pourrait comparer cela à une opération de reconnaissance de formes. Mais il arrive parfois qu'aucune structure significative ne puisse être dégagée.

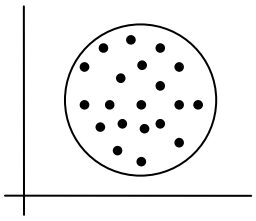
Exemples de structure-type dans un tableau de données :

- Tous les individus de notre tableau ont le même poids et la même taille,

• Les individus se répartissent uniformément dans un quasi-cercle, le fait d'être lourd ne conditionne pas la taille et vice-versa.

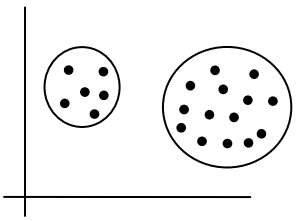


Les 2 variables ne varient pas



Absence de liaison

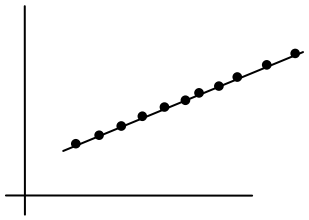
Des formes particulières, dites non linéaires, structurent notre population,



deux sous-nuages

forme parabolique

Les individus se trouvent alignés le long d'une droite, au quel cas, la structure du tableau est unidimensionnelle : lorsque le poids est connu, la taille l'est aussi.



les 2 variables sont parfaitement corrélées

2.3. Exemple à 3 dimensions

Nous avons vu comment on pouvait dégager la structure d'un tableau de données à deux dimensions matérialisé par un nuage de points sur un plan. Que se passe-t-il si l'on ajoute au tableau de données une troisième variable (l'âge des individus par exemple) ?

Notre tableau devient le suivant :

Tableau 2. Exemple de tableau à 3 variables.

Individu	Poids	Taille	Age
A	45	151	15
B	95	197	16
C	56	168	18
D	73	178	19
E	69	171	31
F	84	182	24
G	56	165	19
H	82	172	35
I	67	183	21
J	77	162	25

On peut représenter les individus A, B et E graphiquement dans un espace à 3 dimensions :

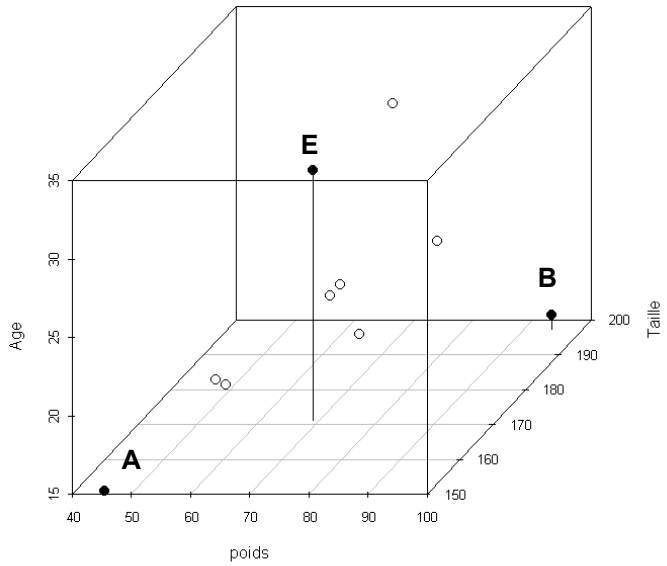


Figure 4. Positions des individus A, B et E dans le nuage à 3 dimensions.

Dans la
Figure 1, nous avons représenté les 3 individus A, B et E qui dans le plan Taille-Poids se trouvaient alignés. Il n'en est plus de même dans le nouveau nuage, l'individu E étant plus âgé (31 ans) que les individus A (15 ans) et B (16 ans). Le nuage comporte une troisième dimension qui modifie sa forme générale.

Il n'est pas très aisé d'examiner un nuage à 3 dimensions. Une solution serait de le réduire en une série de plans, qui eux-mêmes peuvent être interprétés selon leurs axes factoriels, de la même façon que l'on peut résumer un nuage de points en 2 dimensions en deux axes factoriels. On peut aboutir à trouver la structure du nuage tridimensionnel à la lumière de ses 3 axes factoriels.

Dans notre exemple, il semble que la forme des nuages Poids-Age et Taille-Age ne permette pas de dégager une structure forte.

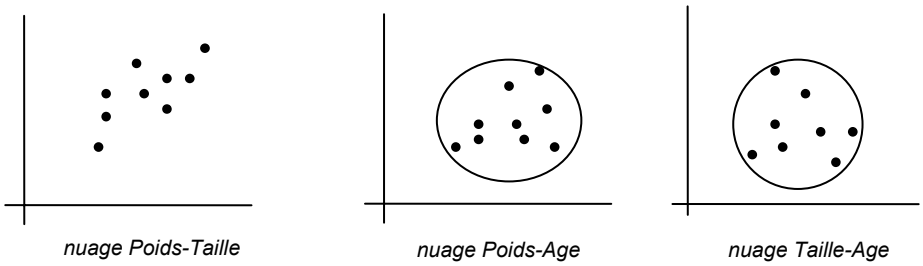


Figure 5. Exploration de la structure du tableau 2 par une série de plans.

En bref, dégager la structure d'un tableau de données,

• Du point de vue intuitif, c'est déceler les dimensions principales, celles qui différencient le plus les individus et éventuellement de trouver les dimensions cachées, l'information secondaire qui n'est pas visible à l'œil nu.

• Du point de vue géométrique, c'est essayer de reconnaître la forme du nuage par quelques traits saillants. Ce sont eux qui caractérisent la nature et l'intensité des relations entre les individus.

• Du point de vue mathématique, c'est la droite ou le plan qui réalise le meilleur ajustement possible du nuage de points.

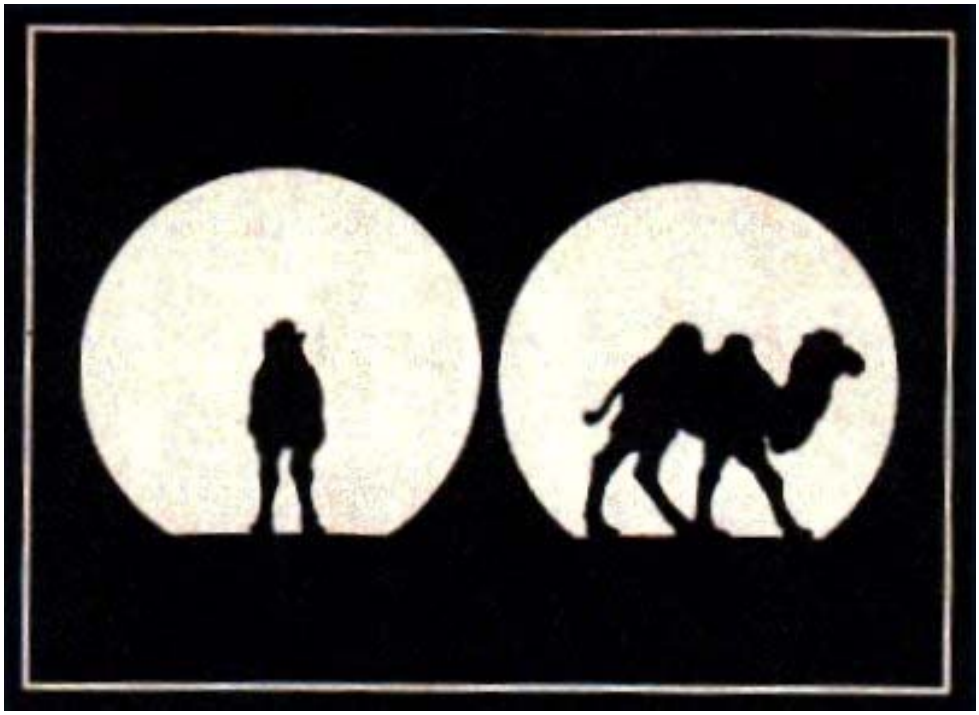


Image 1. L'analyse factorielle propose divers points de vue sur les données.

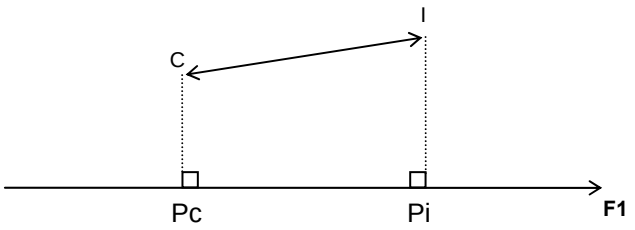
L'analyse factorielle s'apparente à la photographie, c'est à dire représenter une réalité multidimensionnelle sur un support à deux dimensions : la photo. Ceci dit, la photo du chameau et donc l'interprétation de sa forme dépend de l'optique que l'on choisit. Si l'on décide de le regarder de profil, on pourra voir qu'il possède 2 bosses, une queue, quatre pattes, information non discernable vu de face.

2.4. La notion de distance

Lorsque l'on observe un nuage de points, on s'aperçoit aisément que ce sont les proximités ou les éloignements entre les points qui contribuent à donner la forme générale du nuage. Nous avons vu que celle-ci était révélatrice de l'information contenue dans le tableau de données, on peut alors affirmer que ce sont les **distances** entre les individus qui constituent **l'information**. Ainsi, les individus se positionnent dans le nuage au vu des variables qui les décrivent, deux individus proches géométriquement se ressemblent et deux individus éloignés se distinguent ou même s'opposent.

Or lors de l'opération de recherche des axes factoriels, cette information sera quelque peu modifiée. Chacun des points se verra attribuer de nouvelles coordonnées qui tentent de minimiser les écarts entre sa position initiale et celle sur l'axe (Figure 2). Le plus court chemin étant la ligne droite, les nouvelles coordonnées sont déterminées orthogonalement : on appelle cela une opération

de projection orthogonale, opération qui malheureusement a le défaut de réduire les distances entre les points (Figure 6).



Projections orthogonales de C et I sur F1, distance (C,I) ≥ distance (Pc,Pi)

Figure 6. Projection orthogonale sur F1 de la distance entre les points C et I.

De nouvelles distances projetées « déformées » sont donc définies entre les individus. Nous allons voir pourquoi trouver une droite d'ajustement qui minimise les écarts des points avec cette droite est équivalent à maximiser les distances projetées et donc de réduire au mieux la déformation.

Nous allons représenter la distance (C,I) et la distance (Pc, Pi) sous forme d'un triangle rectangle (Figure 7). On peut grâce à Pythagore $\left(d^2 = h^2 + e^2\right)$ facilement calculer *e*. On peut considérer cette quantité comme une mesure de l'écart entre la distance réelle *d* et sa distance projetée *h* sur F1. Pour avoir le moins de perte possible, l'idéal serait que l'axe passe par C et I ou, qu'il soit parallèle à *d*. Donc, réduire e revient à tenter de reproduire au mieux d, soit maximiser *h*.

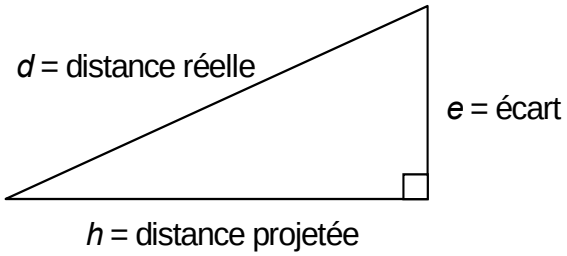


Figure 7. Déformation de la distance entre deux points par la projection orthogonale.

La somme des distances entre les points constitue la dispersion du nuage et l'on tentera de la restituer avec le moins de déformation possible. Tout l'enjeu des méthodes factorielles est là.

3. Les méthodes factorielles d'analyse des données

On suppose que vous ayez fait sur un certain nombre d'individus des observations que vous avez rassemblées dans un nombre important de variables. Il sera très difficile de recenser les relations entre toutes les variables. L'idéal serait de réduire ce nombre de variables à une quantité plus modeste afin de ne retenir que les traits importants. Les méthodes factorielles se proposent de fournir des représentations synthétiques de ces vastes ensembles de données, sous forme graphique dans un espace de faible dimension (en général le plan, qui avec deux dimensions offre la visualisation la plus aisée de nos données, voir Figure 1).

La recherche de cet espace s'effectue par ajustement de deux nuages de points : celui des points-individus qui sont décrits par les variables et celui des points-variables qui sont décrits par les individus, de façon à ce que les proximités mesurées dans ces espaces d'ajustement reflètent autant que possible les proximités réelles. L'espace de représentation que l'on obtient s'appelle l'espace factoriel.

3.1. Différentes méthodes pour différents tableaux de données

Un tableau de données, c'est une matrice de nombres structurée en colonnes et en lignes. Les colonnes peuvent être des **variables quantitatives ou continues** (variables dont les valeurs possibles sont des nombres régies par des relations d'ordre : 3 est plus grand que 2) ou des **variables qualitatives ou nominales** (les valeurs pour chaque variable étant les modalités de réponses à la question-variable). Quant aux lignes, elles peuvent être des individus ou observations ou encore des modalités d'une variable qualitative (cas du tableau de contingence).

La nature des informations, leur codage, fera en sorte que chaque tableau de données sera soumis à une méthode factorielle spécifique. Avant de présenter les méthodes factorielles, faisons donc un inventaire des principaux tableaux susceptibles d'être analysés.

Le tableau de mesures
On a effectué sur un nombre *n* d'individus, *p* mesures que l'on range dans un tableau de données dit : individus-variables. Ce sont des mesures de variables dites quantitatives ou continues qui peuvent s'exprimer dans des unités diverses : taille en cm, nombre de bovins dans un élevage, taux d'accroissement, etc.

Le tableau de contingence
On a classé les individus d'une population selon une première caractéristique, par exemple selon leur activité professionnelle principale et selon une deuxième caractéristique comme les niveaux de revenu en différentes catégories. Le tableau de contingence est celui qui comptabilise les occurrences résultant du croisement des deux caractéristiques. Contrairement aux autres tableaux, les lignes et les colonnes ont des rôles symétriques dont l'étude sera celle de la liaison des deux variables qualitatives.

Tableau 3. Exemple de tableau de contingence.

	Activité professionnelle			
Niveau de revenu	éleveur	agro-éleveur	agriculteur	autre
niveau faible	10	25	15	20
niveau moyen	20	50	50	15
niveau élevé	40	30	20	5

Le tableau qualitatif
On a effectué sur un nombre n d'individus, *p* « mesures qualitatives » que l'on range dans le même type de tableau que le tableau de mesures. A la différence près, que les individus sont décrits par les modalités de variables dites qualitatives. Il peut exister une relation d'ordre mais on ne quantifie pas l'écart entre les modalités. De plus les valeurs des codes sont purement arbitraires. Par exemple, le code 1 pour Sexe désigne « homme » comme on pourrait l'utiliser pour « femme ».

Le tableau présence-absence
C'est en fait un cas particulier de tableau qualitatif où toutes les variables ne comportent que deux modalités : la présence ou l'absence du caractère.
Il faut avoir en tête que dans certains cas, la présence et l'absence ne jouent pas des rôles symétrique. Deux individus qui ont un caractère en commun sont proches, l'inverse pas forcément (relevés écologiques, données génétiques). Il faudra alors veiller à employer des mesures de ressemblances appropriées.

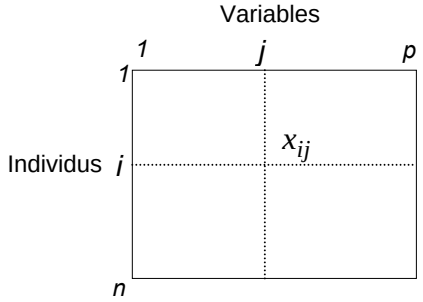
Le tableau de notes
Les codes de note attribués aux individus de notre tableau nous permettent de les ordonner. Ils expriment une intensité et l'on peut alors effectuer un classement de nos individus et assimiler les variables à des descripteurs qualitatifs. L'exemple le plus parlant est celui des notes scolaires. Dans le cas où l'intervalle de variation des valeurs est suffisamment grand, on peut assimiler le tableau de notes à un tableau de mesures quantitatives.

Nous allons décrire ces méthodes une à une en commençant par l'analyse en composantes principales (ACP) dont le schéma général d'analyse est commun aux autres et constituera ainsi, une bonne introduction.

3.2. L'analyse en Composantes Principales (ACP)

3.2.1. Objectifs

C'est la méthode de base pour la recherche de facteurs synthétique de la variabilité d'un ensemble de variables. Cette analyse qui constitue le schéma de calcul de base des analyses factorielles, s'applique aux tableaux de mesures que l'on dénommera X, organisés et notés de la manière suivante :



Le tableau X recense les observations effectuées sur les n individus pour les p variables. x_{ij} est la valeur de la variable *j* pour l'individu *i*.

Les individus et les variables jouent des rôles symétriques mais recouvrent des notions différentes. On cherche pour les individus à évaluer leur ressemblance. Deux individus se ressemblent d'autant plus qu'ils possèdent des valeurs proches pour l'ensemble des variables qui les décrivent. On pourra ainsi par la suite détecter des groupes homogènes d'individus ou certains individus atypiques.

En ce qui concerne les variables, ce sont leurs liaisons qui nous intéressent. Mesurées par le coefficient de corrélation linéaire (noté *r*), on pourra mettre en avant des groupes de variables corrélées ou au contraire des oppositions (liaisons négatives).

Ainsi deux nuages vont être analysés séparément sachant que deux variables seront proches vis à vis des *n* individus et deux individus seront proches vis à vis des *p* variables.

3.2.2. Une transformation du tableau de données

Puisque l'on s'intéresse à la dispersion des points dans un nuage, le calcul des facteurs principaux ne se fait pas à partir du tableau de base mais sur un tableau préalablement **centré**, c'est à dire que l'on a retranché pour chaque valeur x_{ij} du tableau X la moyenne notée $\overline{x}^{j}$ de la variable *j*. C'est donc pour chaque individu, l'écart à la moyenne, plus révélateur de sa position dans le nuage, qui est analysé.

D'autre part, les résultats d'une ACP ne sont pas indifférents à l'importance des quantités inscrites dans les tableaux de données. Ainsi, dans l'exemple Poids-Taille, l'unité choisie pour définir la taille est le cm, et nous aurions eu des résultats différents si nous avions choisi le mètre comme unité.

On comprend mieux à l'aide d'un exemple :
calculons la distance entre l'individu A (45 kg ; 151 cm) et B (95 kg ; 197 cm) :

Grâce à $d^2 = AB^2 = AH^2 + HB^2$ (*Figure 8*),
on a $d^2 = (45-95)^2 + (151-197)^2 = 4616$.

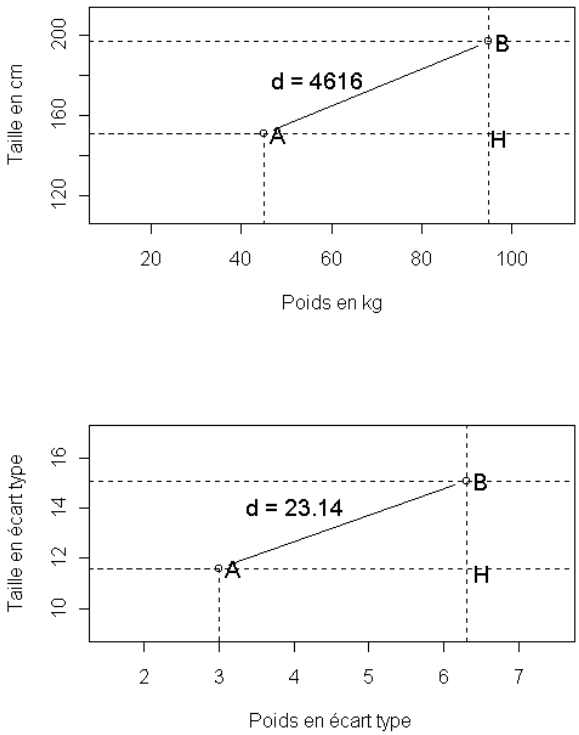


Figure 8. Illustration de la réduction des variables dans le calcul de la distance entre A et B.

En changeant l'unité de la variable Taille de cm en mètre, la distance entre A et B devient : $50^2 + 0.46^2 = 2500$ et où l'on s'aperçoit qu'une variation d'une unité pour le poids a beaucoup plus d'importance que la même variation sur la taille du fait des unités utilisées et ce d'autant plus que les distances sont calculées au carré.

Afin d'éliminer ce biais, la solution est d'utiliser une unité commune : l'écart-type. En **réduisant** les données, c'est à dire en divisant par l'écart-type de la variable correspondante, on donne la même importance aux variables pour le calcul des distances entre individus.
Ainsi, l'écart-type (noté σ) de la distribution des poids est 15 kg. La longueur AH représente donc 50/15 = 3.3σ. De même pour la distribution des tailles : σ=13, ainsi HB = 46/13 = 3.5σ.

La distance entre l'individu A et l'individu B est de $3.3^2 + 3.5^2 = 23.14$

Si l'on fait une ACP sur les données du tableau X qui a subi une transformation du type : retrancher la moyenne de la variable et diviser par l'écart-type, on dit que l'on fait une ACP sur données centrées-réduites ou **ACP normée**. C'est la méthode d'ACP la plus couramment utilisée car on dispose la plupart du temps de données hétérogènes (unités de mesures différentes pour les variables).

3.2.3. Analyse du nuage des individus

Tous les individus du tableau X sont matérialisés par des points dans un espace à autant de dimensions qu'il existe de variables qui les caractérisent. Afin de déchiffrer la forme du nuage pour en saisir l'information, l'ACP procède à un ajustement : elle recherche un premier axe qui traverse le nuage dans son allongement le plus grand afin de minimiser les distances entre les points et leur

projection orthogonale sur l'axe ou ce qui est équivalent (du fait du centrage), de maximiser les distances entre tous les couples de points projetés sur cet axe.

La somme des distances entre tous les points et leur **centre de gravité** (point moyen général ou individu moyen, noté *G*) est la mesure de la dispersion du nuage de points que l'on appelle l'**inertie** et qui est équivalente à la variance mais dans un espace multidimensionnel.

Ainsi la procédure de calcul recherche la direction pour l'axe maximisant cette inertie qui est projetée sur celui-ci (Figure 9) pour essayer de restituer au mieux « l'allongement » du nuage initial qui correspond à la source de variabilité principale entre les individus. Cette recherche d'axes se fait séquentiellement. Le premier axe est le plus long d'allongement du nuage, ainsi de suite pour les autres qui ont pour contrainte d'être orthogonaux et de passer par *G*.

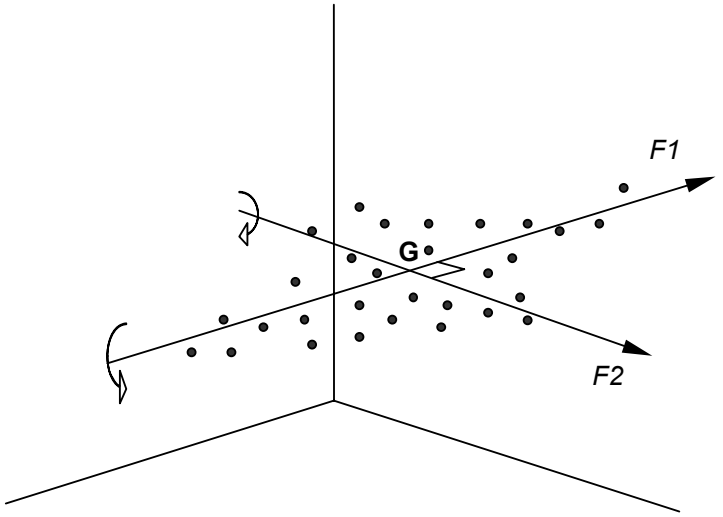


Figure 9. Recherche des directions principales du nuage par les axes factoriels.

Les individus se voient alors attribués de nouvelles **coordonnées** dites **factorielles** pour tous les axes calculés. Une image synthétique ou résumée du nuage nous est donnée par le plan formé par les deux premiers axes de **l'espace factoriel** (Figure 10).

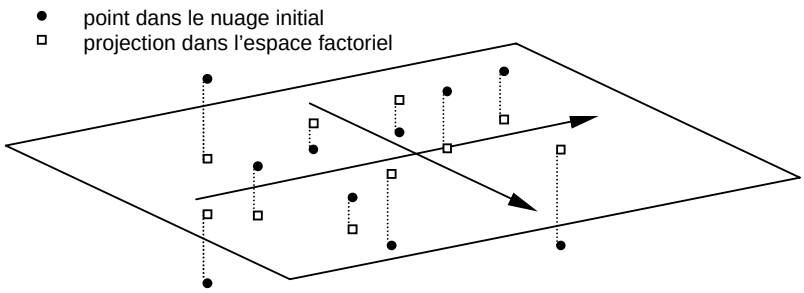


Figure 10. Projections des points sur l'espace factoriel F1-F2.

Les premiers axes (les premiers plans) relatent les proximités (ressemblances) et éloignements (dissemblances) **principaux** entre nos individus.

Ces **axes factoriels** sont ce que l'on appelle les **composantes principales** (notées F1, F2, …) du tableau de données. Orthogonaux, non corrélées, elles apportent chacune une information indépendante. Dans des termes plus

mathématiques, ce sont de nouvelles variables, combinaisons linéaires des variables initiales, de la même manière que dans le premier exemple où l'axe factoriel 1 exprimait la corpulence de nos individus par une combinaison de la taille et du poids.

3.2.4. Analyse du nuage des variables

L'ACP applique au nuage des variables la même démarche que pour le nuage des individus sauf que l'espace de représentation des variables est différent.

Chacune des variables est décrite par une colonne du tableau X (les valeurs qu'elle prend pour les *n* individus) et peut donc être représentée graphiquement dans un espace à autant de dimensions qu'il y a d'individus. Les points-variables sont inscrits dans une sphère de centre 0 et de rayon 1 (Figure 11) du fait de la normalisation et de même que pour les individus, il existe une distance entre chacun d'entre eux qui n'est autre que la taille de l'angle formé par les vecteurs supports. Cette distance se mesure en prenant le cosinus carré de l'angle qui, grâce au fait que nos variables sont centrées-réduites, est égale au coefficient de corrélation linéaire bien connu pour mesurer la liaison entre deux variable. Un petit angle signifiera un $\cos^2$ proche de 1 (de même par symétrie, un grand angle correspond à un $\cos^2$ proche de -1) donc une bonne corrélation.

Les distances entre les variables s'appréhendent en terme de corrélation. C'est le coefficient de corrélation linéaire qui est l'indicateur de forme du nuage de points variables, la particularité principale de ce nuage par rapport à celui des individus est là.

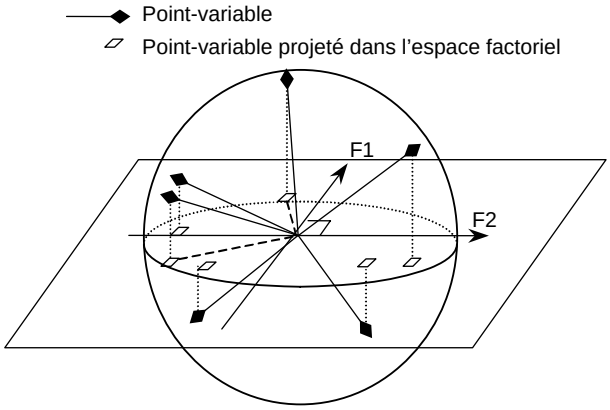


Figure 11. Représentation graphique des variables normées.

Les distances entre les points-variables sont difficilement discernables au sein de la sphère, il s'agit donc de trouver un espace de représentation de dimension moindre qui nous permette de reproduire les angles le plus fidèlement possible.

On ajuste dans la sphère, un axe sur lequel on projette orthogonalement les points-variables. On l'a vu pour les individus, reproduire un nuage de points revient à « récupérer » le maximum de sa dispersion. Le critère d'ajustement sera donc celui de maximisation de l'étalement du nuage (inertie) projeté, ce qui pour notre nuage, revient à maximiser la somme des $\cos^2$ des angles α formés par les points-variables et leur projection (Figure 12). Ces mesures étant égales au coefficient de corrélation linéaire entre la variable et l'axe factoriel, on recherche ainsi un axe synthétique F1 **le plus lié** à l'ensemble des variables. Et ainsi de suite pour les axes factoriels suivants orthogonaux, c'est à dire non corrélées entre eux essayant de restituer au mieux les liaisons résiduelles.

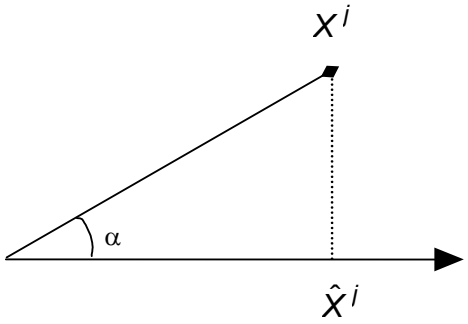


Figure 12. Projection orthogonale du point-variable X^i sur le facteur.

3.2.5.La dualité

Ainsi, comme on l'a vu les nuages des individus et des variables sont des représentations de la même information, du même tableau de données. Des relations dites de dualité lient les deux nuages qui sont assez faciles à comprendre.

En effet, prenons l'axe factoriel F1 synthétisant la corrélation d'un groupe de trois variables. Si elles le sont, c'est par le fait d'individus qui prennent des valeurs fortes ou faibles pour ces trois variables. Ainsi cet axe original pour ce groupe de variables l'est pour le groupe d'individus. Et il en va de manière symétrique pour les individus par rapport aux variables. Ces relations, on va le voir, sont primordiales pour l'interprétation même si les plans factoriels respectifs ne sont pas superposables.

3.2.6. Aides à l'interprétation

L'exposé détaillé du calcul des axes factoriels va nous permettre de passer à l'étape essentielle pour l'utilisateur des méthodes factorielles : l'interprétation.

Le résultat principale d'une ACP ou de tout autre méthode factorielle est une image de notre tableau de données sous forme graphique : **le plan factoriel**. Or cette image est un peu déformée par le traitement dont elle est le résultat. Afin de lire correctement cette image, il existe des indices numériques accompagnant l'image factorielle que l'on appelle les aides à l'interprétation. On fera ici un bref descriptif afin de se familiariser avec ces indicateurs et l'on y reviendra dans notre exemple des éleveurs du Breedland.

« Interpréter, c'est donner un sens aux facteurs », procèdant axe par axe, en tous les cas pour ceux que l'on juge significatif du point de vue de l'information recueillie. L'examen du plan factoriel des variables permet de visualiser les corrélations et celui des individus d'identifier des groupes d'individus ayant pris les mêmes valeurs pour les mêmes variables. Mais ces axes restaurent une partie seulement de la dispersion initiale. On a donc à disposition des outils permettant d'évaluer la qualité de l'approximation[1].

Le diagramme des valeurs propres

Les valeurs propres sont un des produits du calcul des axes factoriels. Chacune s'associe à un axe pour donner une mesure de la dispersion (inertie) qui a été

[1] Nous sommes toujours dans le cadre de l'interprétation d'une ACP normée.

projetée sur celui-ci. Ou formellement, la valeur propre associée à un facteur est la dispersion des coordonnées factorielles des points-individus projetés.
On se ramène le plus souvent au pourcentage d'inertie expliquée cumulée par plusieurs facteurs, si celui-ci est proche de 100% pour un petit nombre de facteurs, on aura reconstituer entièrement le nuage initial[2] à l'aide d'un espace de faible dimension.

Une pratique courante est de retenir un nombre d'axe correspondant à un nombre de barres du diagramme au delà desquels on observe une rupture Figure 13). Cette rupture est significative de la quantité d'inertie (information) qui a été projetée sur les facteurs retenus. Avec plus de deux axes, on fait le dépouillement par plans successifs (F1-F2, F1-F3, F3-F4, etc.) en choisissant les couples d'axes où les points sont le mieux représentés.

Si le diagramme indique une décroissance régulière de l'inertie projetée, c'est qu'il n'y a pas de structure forte dans le tableau de données. Dans ce cas précis, c'est l'homogénéité des individus et des corrélations faibles entre variables qui sont mises en évidence par l'ACP.

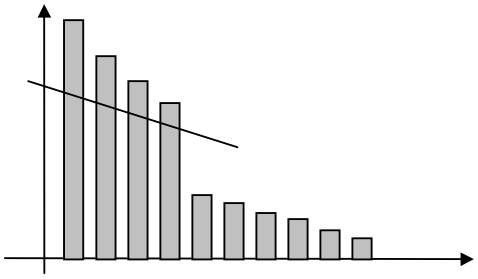


Figure 13. Diagramme des valeurs propres.

La contribution d'un point à l'inertie projetée sur un axe

Pour les individus,
L'individu est détenteur d'une certaine quantité d'informations, qui est en réalité sa distance par rapport au centre de gravité du nuage. Un point-individu éloigné sera prépondérant dans la forme générale du nuage et donc dans l'orientation de l'axe qui va ajuster ce nuage (Figure 14). On peut mesurer cette importance grâce à l'indicateur **contribution** qui est le rapport *dispersion projetée de l'individu/dispersion totale du nuage projetée* et permet de savoir quels sont les points qui ont participé en premier lieu à la construction d'un axe et sur lesquels s'appuiera l'interprétation.

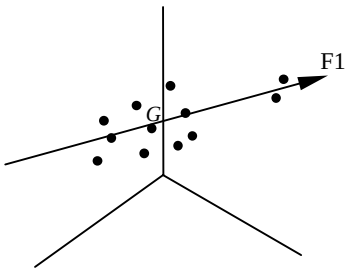


Figure 14. Contribution d'un point à l'inertie du nuage.

Pour les variables,

[2] On parle ici indifféremment de l'inertie du nuage des individus et de celui des variables, car du fait de la dualité, elle sont égales.

En ACP normée, l'inertie se lit en terme de corrélation, ainsi les variables fortement corrélées avec un axe vont contribuer à la définition de cet axe. La corrélation se lit directement : il s'agit de l'angle formé par la variable avec l'axe factoriel (Figure 12). Lors de l'examen de ces corrélations, il peut arriver qu'un nombre important de variables soient corrélées positivement avec un axe (généralement le premier). On peut le voir facilement à l'aide du cercle des corrélations figurant sur le plan factoriel (Figure 15). On a mis en évidence ce que l'on appelle en ACP, **un effet taille**. Ceci signifie simplement que ces variables varient dans le même sens et que le facteur concerné classe les individus depuis ceux qui présentent les plus faibles valeurs pour ces variables jusqu'à ceux qui présentent les plus fortes. Dans ce cas de figure, nos variables sont proches (du point de vue géométrique), détentrices de la même information (du point de vue statistiquement) puisque corrélées (du point de vue mathématique).

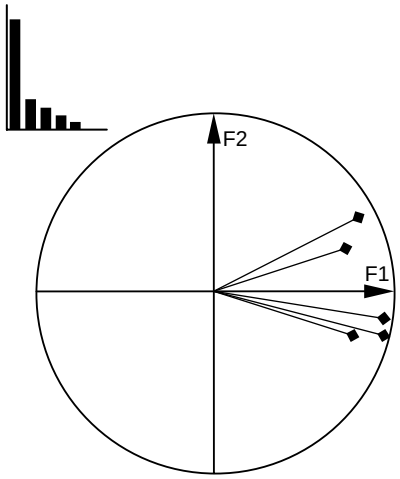


Figure 15. L'effet taille en ACP.

La qualité de représentation des points

L'image factorielle d'un nuage de points est forcément déformée, elle induit notamment des erreurs de perspective puisque tous les points positionnés dans un espace de dimension élevée se voient figurer dans un espace de visualisation à seulement deux dimensions. **La qualité de représentation** des points projetés mesurée par le **$\cos^2$** nous permet de savoir quels sont les points mal représentés afin de mieux évaluer leurs proximités relatives sur les axes et les plans factoriels. La qualité de représentation d'un point est répartie et elle peut faire l'objet d'une sommation axe par axe.

Pour les individus,
Prenons deux points M et N tels que dans leur nuage, ils se situent l'un au dessus de l'autre par rapport au premier axe factoriel (Figure 16) . Ces deux points se projettent au même endroit sur l'axe F1, pourtant ils sont en réalité éloignés l'un de l'autre.
Afin de mesurer cette déformation liée à l'ajustement, on peut prendre le $\cos^2$ de l'angle formé par les droites qui lient les points M et N et l'axe factoriel. Dans le cas présent, le $\cos^2$ ($\alpha 1$) est inférieur au $\cos^2$ ($\alpha 2$) nous permettant de relativiser les proximités de M et N sur F1. Cet indice nous permet également d'évaluer la distance d'un point au centre de gravité G.
Dans un cas extrême, l'individu Z se projette aux environs du centre de gravité mais son $\cos^2$ proche de 0 nous met en garde contre cette déformation nous indiquant l'éloignement de Z à G dans le nuage initial.

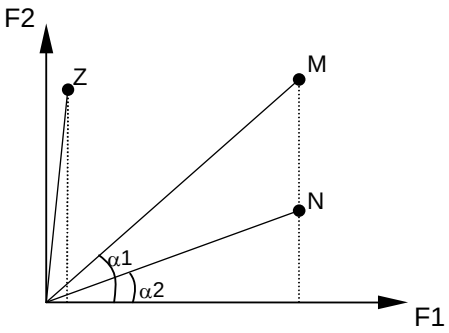


Figure 16. Mesure de la qualité de la représentation des points M et N par le cos² de l'angle α.

Autre point qui a son importance, ces mesures de qualités de représentation sont additives par axe. On peut donc évaluer la qualité de représentation d'un individu sur un plan.

Pour les variables,
En pratique, on trace dans l'espace factoriel des variables un cercle de rayon 1 et d'origine 0 dit de corrélation qui permet de juger de la qualité de représentation des variables. Un point-variable sera bien projeté si sa coordonnée factorielle (qui est aussi le coefficient de corrélation avec l'axe uniquement dans le cadre d'une ACP normée) est proche de 1, c'est à dire proche du **cercle de rayon 1 dit des corrélations**. En effet, les variables initialement le plus près de l'axe d'ajustement (Figure 11) l'orientent dans leur direction et voient leur projection sur celui-ci la moins déformée.
On se gardera donc d'interpréter des proximités (corrélations) entre points-variables si ceux-ci ne sont pas proches du cercle des corrélations.

Les points supplémentaires

La pratique des points supplémentaires ou illustratifs fait partie des aides à l'interprétation. Très utilisée, elle permet de juxtaposer sans participer activement aux calculs des axes, une information complémentaire afin d'éclairer les résultats d'une analyse du tableau de données dit **tableau actif**.

Il est bien entendu que ces points supplémentaires ont une contribution nulle à la dispersion projetée mais leur position par rapport aux facteurs se lit après avoir vérifié la qualité de leur projection à l'aide du cos². Les points supplémentaires peuvent être :

Des variables,

- Celles-ci ne présentent pas le même intérêt que celles du tableau actif, elles peuvent prendre le rôle de variables « à expliquer ». Par exemple les opinions des éleveurs sur les problèmes de développement pourront être mis en supplémentaires afin d'être expliqué par les variables actives de pratiques d'élevage.

- Afin de mélanger des variables dont la thématique est différente comme la structure des élevages en variables actives et les pratiques, ou les pathologies que l'on a pu observer en variables supplémentaires.

- Pour écarter une variable trop contributive du fait de modalités rares, de valeurs manquantes trop nombreuses, ou trop prépondérantes dans la différenciation des individus (par exemple l'ethnie des éleveurs ou la taille de

l'exploitation qui servent dans certains cas de variable de stratification pour l'échantillonnage).

Des individus,

- Que l'on positionne à posteriori[3] pour juger de leur similitude avec les individus actifs. La technique des individus supplémentaires peut servir aussi à « éclater » le nuage de points en isolant des individus qui perturbent fortement la détermination des axes factoriels). Par exemple, on met en supplémentaires des éleveurs trop atypiques qui gênent l'analyse, celle-ci mettant uniquement en relief l'opposition avec les autres éleveurs et masque les oppositions existantes dans ce groupe.

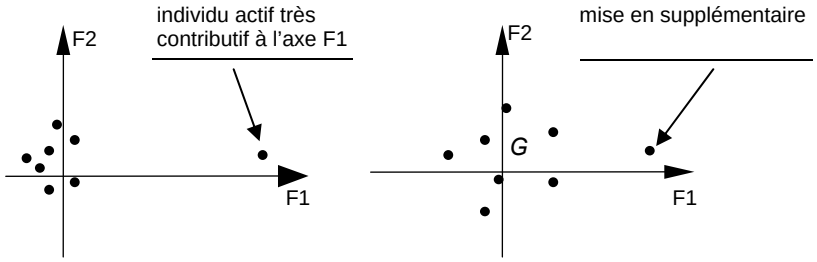


Figure 17. Individu très contributif mis en supplémentaire.

3.2.7. Exemple : caractéristiques pondérales des carcasses de bovins

Pour terminer cette présentation de l'ACP dont les outils et les modes d'interprétation sont centraux en analyse factorielle, nous allons faire l'étude des caractéristiques pondérales de deux échantillons de deux races de bovins (charolais et zebus).

On active la librairie ade4 après s'être assuré qu'elle a bien été installée (se reporter en fin de document au chapitre 'Outils logiciels' pour des précisions sur cette librairie) :

```
> library(ade4)
```

Importer le tableau dans **R** :

```
> zebus <- read.table("zebus.txt", sep=";", header=T)
```

Afficher quelques statistiques descriptives :

```
> summary(zebus)
      VIF             CARCA           V1QUAL         VTOTALE
 Min.   :390.0   Min.   :213.0   Min.   :25.80   Min.   :70.30
 1st Qu.:395.0   1st Qu.:223.0   1st Qu.:28.00   1st Qu.:72.95
 Median :400.0   Median :229.0   Median :30.40   Median :74.50
 Mean   :401.2   Mean   :228.8   Mean   :29.92   Mean   :74.67
 3rd Qu.:405.0   3rd Qu.:234.0   3rd Qu.:31.90   3rd Qu.:76.50
 Max.   :420.0   Max.   :247.0   Max.   :35.10   Max.   :79.10
      GRAS             OS              RACE
 Min.   : 4.600   Min.   :14.50   charolais:12
 1st Qu.: 7.350   1st Qu.:15.60   zebu     :11
 Median : 8.700   Median :16.40
 Mean   : 8.974   Mean   :16.42
 3rd Qu.:10.700   3rd Qu.:17.15
 Max.   :13.500   Max.   :18.70
```

Faire l'ACP normée du tableau zebus en omettant le facteur RACE :

```
> res <- dudi.pca(zebus[,-7], center=T, scale=T)
```

[3] Après la construction des axes factoriels.

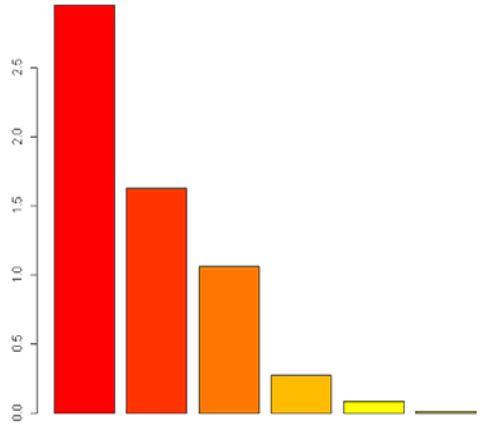


Figure 18. Diagramme des valeurs propres de l'ACP normée du tableau zebus.

Si l'option `scannf=T`, la fonction affiche le diagramme des valeurs propres et attend une saisie du nombre de facteurs :

```
Select the number of axes: 3
```

Tous les résultats de l'ACP sont rangés dans l'objet de type liste `res` :

```
> res
Duality diagramm
class: pca dudi
$call: dudi.pca(df = zebus[, -7], center = T, scale = T)

$nf: 3 axis-components saved
$rank: 6
eigen values: 2.951 1.625 1.061 0.2715 0.08103 ...
  vector length mode    content
1 $cw    6      numeric column weights
2 $lw    23     numeric row weights
3 $eig   6      numeric eigen values

  data.frame nrow ncol content
1 $tab       23   6    modified array
2 $li        23   3    row coordinates
3 $l1        23   3    row normed scores
4 $co        6    3    column coordinates
5 $c1        6    3    column normed scores
other elements: cent norm
```

L'ensemble des éléments utiles pour l'interprétation mais également à des opérations ultérieures sont disponibles. Le tableau des données transformées, dans notre cas normalisées, est accessible grâce à la commande :

```
> res$tab
          VIF       CARCA     V1QUAL     VTOTALE       GRAS          OS
1  -0.8220599 -0.56991313  2.0592412  1.82769758 -1.2626805 -1.37577936
2   1.1751984  0.37480773  0.7866958 -0.52373670  0.3082860 -0.01965399
3   0.5094456  0.49289784  0.3094913  0.75511352 -0.6258022  0.07075437
4   0.5094456  1.31952859  0.1901902  0.26007473 -0.1162995 -0.38128742
5  -1.4878127 -1.39654387  0.7866958  0.75511352 -0.4984265 -0.65251250
6   0.5094456  1.67379891  0.8662299  1.12639262 -0.7956364 -0.83332921
7  -1.4878127  0.02053741  0.8662299  1.53892495 -1.8571003  0.52279616
8   0.5094456  1.31952859  0.4685595  0.75511352 -0.3285923 -1.01414593
9   2.5067039  0.61098794  0.9855310  0.54884736 -0.7531779  0.34197944
10 -1.4878127 -0.68800323  1.5422696  0.96137969 -1.1777634  0.34197944
11  1.8409512  2.14615934  0.3094913  0.34258119 -0.2436752 -0.29087906
12 -0.1563072  0.61098794  0.7071617  1.20889909 -1.3900562  2.05973824
13 -0.1563072 -0.56991313 -0.6846847 -0.48248346  0.8602473 -0.83332921
14 -0.8220599  0.02053741 -0.2074802 -0.06995113  0.1384518 -0.29087906
15 -0.8220599 -1.16036366 -0.0881791 -0.77125609 -0.1162995  1.87892153
16 -0.8220599 -0.56991313 -0.5256165 -0.39997700 -0.1162995  0.79402123
17 -0.1563072 -0.68800323 -0.5653836 -0.64749639  0.0535347  1.15565466
```

```
18 -0.1563072 -0.56991313 -0.8437529 -0.60624316  1.3697499 -1.64700444
19 -0.1563072 -0.92418345 -1.3607244 -0.97752226  1.7943355 -1.73741279
20  1.1751984  0.49289784 -1.5993267 -0.97752226  0.9027058  0.16116273
21  0.1099939  0.61098794 -1.1221222 -1.06002872  0.6054959  0.97483795
22 -0.1563072 -0.68800323 -1.2414233 -1.80258692  1.9217112 -0.20047071
23 -0.1563072 -1.86890430 -1.6390937 -1.76133368  1.3272914  0.97483795
```

Le résultat essentiel, les coordonnées factorielles des individus et des variables sont rangées dans la liste `res` sous deux `data.frame` distincts :

```
> is.data.frame(res$li)
[1] TRUE

> is.data.frame(res$co)
[1] TRUE
```

Effectuer des calculs sur certains de ces éléments est tout à fait possible. Par exemple, la distribution en pourcentage de la variance projetée sur les facteurs s'obtient :

```
> round(res$eig/sum(res$eig)*100, 2)
[1] 49.18 27.09 17.68  4.52  1.35  0.18
```

La fonction graphique générique `scatter (scatter.dudi)` permet un affichage condensé des résultats graphiques (Figure 19) :

```
> scatter(res)
```

Préciser les axes si besoin (Figure 20) :

```
> scatter(res, 1, 3)
```

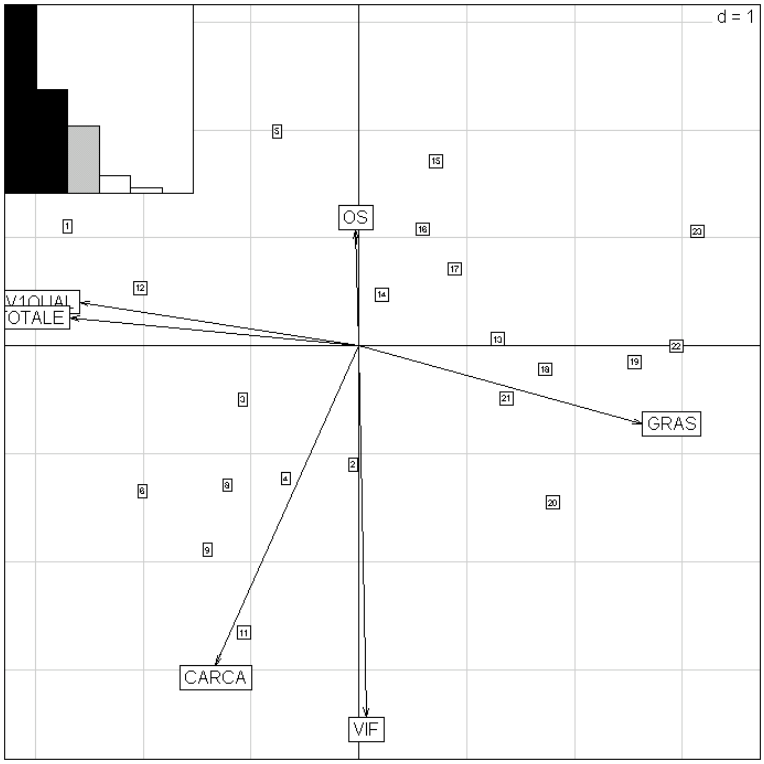


Figure 19. Représentation des bovins et des caractéristiques pondérales sur le plan factoriel 1-2.

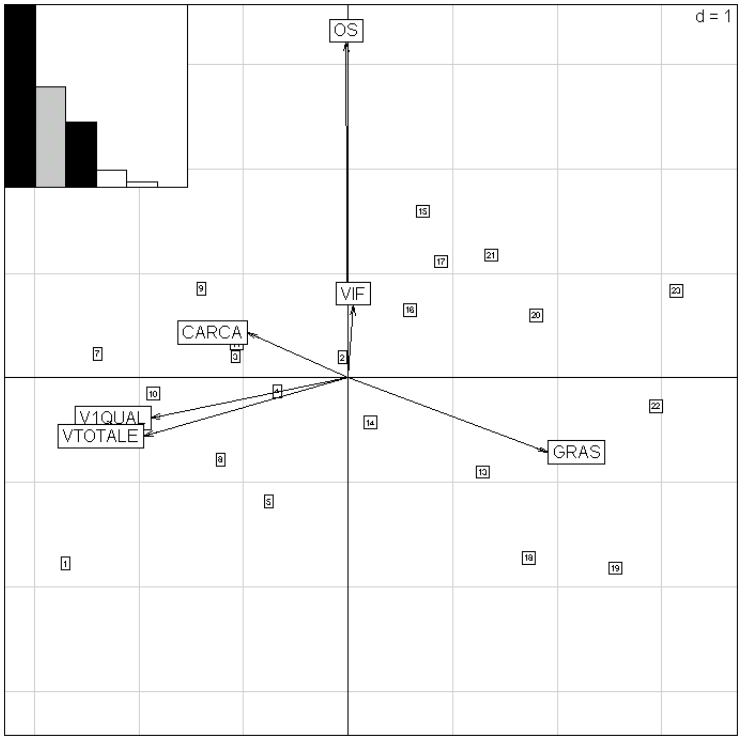


Figure 20. Représentation des bovins et des caractéristiques pondérales sur le plan 1-3.

Les statistiques d'inertie ou aides numériques à l'interprétation sont calculées à l'aide de la fonction `inertia.dudi`. Sauvegarde des statistiques d'inertie pour les variables uniquement :

```
> aides <- inertia.dudi(res, col.inertia=T)
```

Afficher les contributions exprimées pour dix mille (contributions absolues dans la terminologie ade4) des variables aux facteurs ou composantes (Comp) sélectionnés :

```
> aides$col.abs
        Comp1 Comp2 Comp3
VIF         2  5255   403
CARCA     781  3896   156
V1QUAL   2972    73   126
VTOTALE  3184    31   258
GRAS     3061   233   427
OS          0   512  8630
```

Les qualités de représentation des variables sur les facteurs (contributions relatives). Le signe est celui de la coordonnée factorielle de la variable sur le facteur :

```
> aides$col.rel
        Comp1 Comp2 Comp3 con.tra
VIF         6 -8541   427    1667
CARCA   -2303 -6333   165    1667
V1QUAL  -8770   119  -134    1667
VTOTALE -9394    50  -274    1667
GRAS     9032  -379  -453    1667
OS         -1   833  9153    1667
```

La librairie ade4 dispose également d'outils qui permettent d'étudier graphiquement les liens entre variables initiales et facteurs, par exemple pour le premier facteur :

```
> score(res, xax = 1)
```

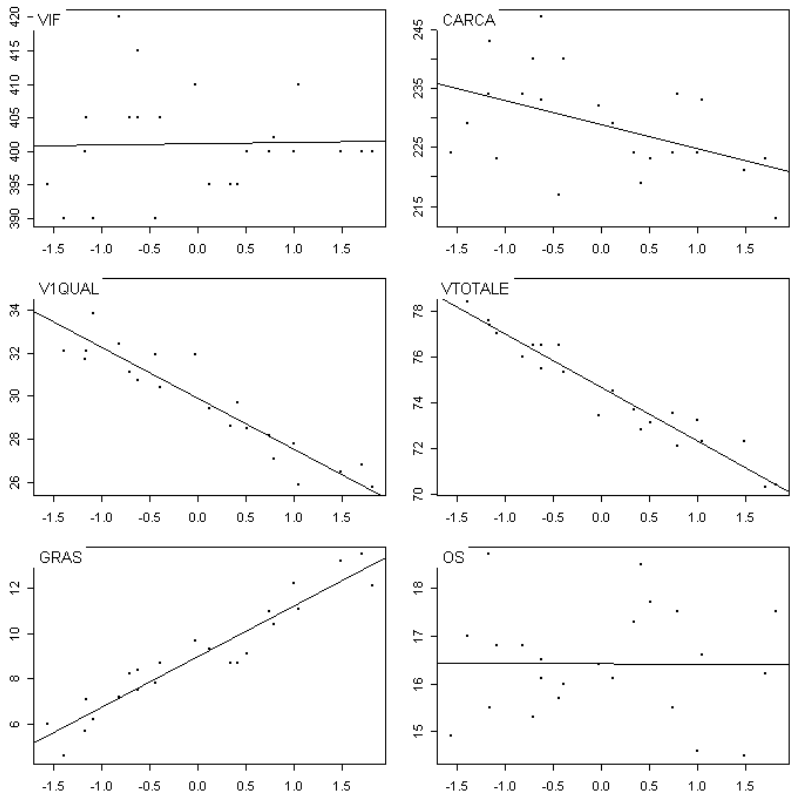


Figure 21. Diagrammes de dispersion facteur 1 (en abscisse) et variables initiales (en ordonnée).

L'étude des liens entre facteurs et une variable qualitative peut s'avérer particulièrement utile pour l'interprétation des résultats d'une analyse factorielle. `s.class` positionne les centres de classes de la variable race et propose un mode de représentation de la dispersion autour de ces centres :

```
> s.class(res$li, zebus$RACE, xax=1, yax=2, cstar=1, cellipse=1.96)
```

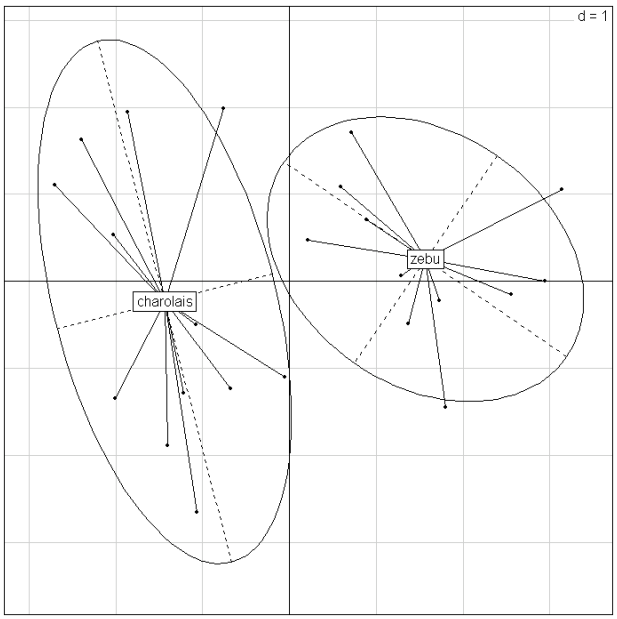


Figure 22. Représentation du lien entre race et les facteurs 1 et 2 de l'ACP de Zebus.

Modifier les options `cstar` et `cellipse` afin de conserver ou pas l'une ou l'autre des représentation de la dispersion des individus autour de leur centre.

3.2.8. En résumé

L'ACP est une méthode descriptive qui met en évidence graphiquement l'information principale contenue dans un tableau de données quantitatives. Comme toute les méthodes factorielles, l'ACP s'appuie sur les distances entre points (variables ou individus) pour synthétiser la dispersion et rendre compte de la structure des données mais aussi des données peu conformes ou aberrantes.

Cette dispersion se traduit pour les variables en corrélation et pour les individus en similitude vis à vis des variables qui les décrivent, ce qui dans le premier cas permet de comprendre en quoi les individus se distinguent et dans le deuxième quels sont les individus qui se ressemblent ou s'opposent.

4. L'analyse factorielle des correspondances

L'analyse factorielle des correspondances permet d'étudier des tableaux dits de contingence. Ce sont des tableaux de comptage obtenus en croisant les modalités de deux variables qualitatives définies sur une même population d'individus. **L'analyse factorielle des correspondances** (**AFC**) est la méthode qui permet de dégager les éventuelles liaisons, dépendances, correspondances existantes entre les variables du tableau.

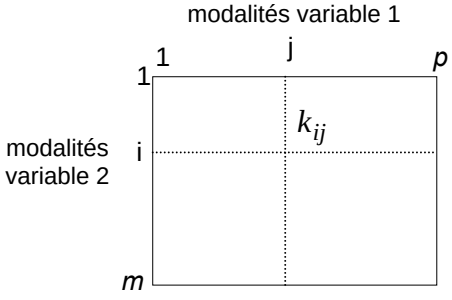
On verra par la suite, que l'on peut généraliser cette analyse à un nombre de variables supérieur à deux moyennant un codage particulier du tableau. Cette généralisation constitue **l'analyse factorielle des correspondances multiples** (**AFCM**) très utilisée pour l'étude des tableaux construits à partir de questionnaires issus d'une enquête.

On n'oublie pas que parallèlement à la décomposition des liaisons entre deux ou plusieurs variables, l'analyse des correspondances est une analyse factorielle. A savoir, une méthode qui cherche à réduire le nombre de dimensions initiales afin de dégager l'information essentielle. Mais commençons par le plus « simple », le cas de deux variables qualitatives.

4.1. Le tableau de contingence

La présentation des notions de l'AFC s'appuiera sur un exemple simple d'étude d'un tableau de contingence.

Soit le tableau de contingence K, k_{ij} recense le nombre d'individus possédant à la fois la modalité *i* de la première variable et la modalité *j* de la seconde.



Soit par exemple le tableau[4] `regicu` qui ventile les réponses des agriculteurs selon leur région d'appartenance et le ou les produits qu'ils cultivent.
Le tableau de contingence est rangé dans un `data.frame` :

```
> regicu <- read.table("regicu.txt", sep=";", header=T, row.names=1)
> regicu
      manioc mais mil riz cafe tabac
nord      10  108 120 425   20    74
sud      388  360 269  40   81    62
est       51   99 150  91  410   342
ouest    209  180 164  80  178   110
```

L'objectif principal de l'étude d'un tel tableau par l'AFC est de savoir s'il y a **indépendance** entre les productions agricoles et les régions d'appartenance des agriculteurs, en d'autres termes, existe-t-il des concordances dans les réponses de nos individus, des **associations** privilégiées entre régions et produits ?

4.2. Transformation du tableau de données

Très simplement, on lit dans ce tableau que 388 individus qui produisent du manioc sont localisés dans la région Sud et que 425 producteurs de riz sont au Nord, mais ce qui nous intéresse, c'est la répartition des cultures dans les régions et l'implantation géographique des cultures et qui ne peut s'appréhender que si l'on considère les **distributions lignes** et **colonnes** respectives.

Ce sont ces distributions qui vont être comparées et pour ce faire, nous allons, de la même façon que dans l'ACP calculer des distances entre les lignes d'une part et les colonnes d'autre part. Ces distances risquent, si on ne rapporte pas chacun des termes (fréquences relatives) à la fréquence marginale du tableau correspondant, d'être fonction du poids des régions et du poids des cultures, c'est à dire du nombre de réponses dans les cellules marginales (totaux lignes et colonnes).

Donc pour comparer efficacement les distributions des lignes et des colonnes, nous allons travailler sur deux tableaux distincts mais symétriques : celui des **profils lignes** et celui des **profils colonnes**, qui sont les répartitions en pourcentage à l'intérieur des lignes et des colonnes.

L'AFC étudie donc les « formes » plutôt que les « tailles », ce qui constitue un avantage énorme. Le fait de traiter des **profils** nous permet d'agréger des lignes et des colonnes semblables sans pour autant changer les résultats. Si par exemple, on estime que l'Est et l'Ouest sont deux régions semblables (au sens où nos agriculteurs y cultivent les mêmes produits), on pourra les réunir en une seule région. Cette propriété très utile dans la pratique est dite d'**équivalence distributionnelle**.

[4] remarquez que les individus de ce tableau de contingence sont des modalités, les lignes et les colonnes du tableau soumis à l'AFC sont de même nature.

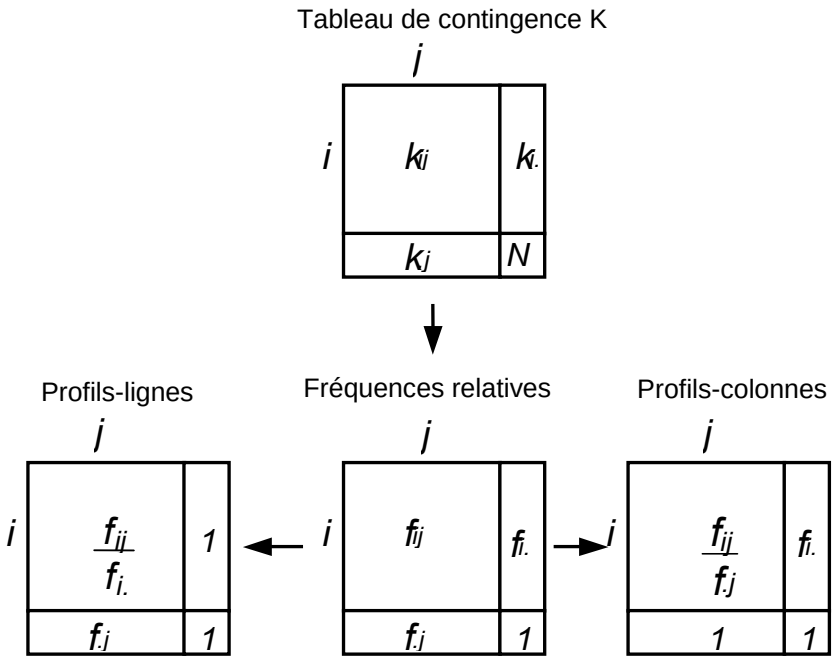


Figure 23. Transformations du tableau de contingence K.

La transformation du tableau K est représentée par les notations suivantes (Équation 1, d'après [2]) :

Effectif total

$$N = \sum_{i=1}^{m}\sum_{j=1}^{p} k_{ij}$$

Fréquences brutes lignes

$$k_{i.} = \sum_{j=1}^{p} k_{ij}$$

Fréquences brutes colonnes

$$k_{.j} = \sum_{i=1}^{m} k_{ij}$$

Fréquence relative cellule i,j

$$f_{ij} = \frac{k_{ij}}{N}$$

Fréquences relatives lignes

$$f_{i.} = \frac{k_{i.}}{N} = \sum_{j=1}^{p} f_{ij}$$

Fréquences relatives colonnes

$$f_{.j} = \frac{k_{.j}}{N} = \sum_{i=1}^{m} f_{ij}$$

Équation 1. Notations des distributions de fréquences.

4.3. La ressemblance entre profils

On n'oublie pas l'objectif fondamental d'une AFC qui est de mettre en évidence des liens éventuels entre les modalités de deux variables. Il faut trouver une mesure qui permette de juger de ces dépendances, on va voir que ces dernières se ramènent à des mesures de distance que l'on nomme **distance du *khi2*** et notée également χ^2.

Dans le tableau régions-cultures, on définit le **profil moyen ligne ou profil marginal** $f_{.j}$ qui est la répartition des cultures quelque soit la région, ce qui correspond bien à une répartition des individus sous l'hypothèse que la production agricole n'est pas fonction de la localisation géographique.

Se référer au profil marginal revient à considérer les profils-lignes identiques entre eux et conclure alors à l'inexistence de relations entre les deux variables. Pour vérifier cette hypothèse, on calcule les distances entre les couples de profils :

Pour mesurer la distance entre deux profils lignes *i* et *i'*, on utilise la formule suivante :

$$d^2(i,i') = \sum_{j=1}^{p} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2$$

Équation 2. Distances entre 2 profils lignes.

Remarquons que l'on a pondéré cette distance par l'importance relative de la culture dans chaque région ($\frac{1}{f_{.j}}$ = l'inverse de la fréquence de chaque colonne), ceci afin de ne pas privilégier une culture dominante du point de vue absolu. Ceci dit, cette pondération a pour effet de donner plus de poids lors du calcul de la distance entre profils régionaux, à la région dans laquelle est cultivé un produit rare.

De même entre deux profils colonnes *j* et *j'*, la formule s'écrit :

$$d^2(j,j') = \sum_{i=1}^{m} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2$$

Équation 3. Distance entre 2 profils colonnes.

avec la pondération : inverse de l'importance du nombre d'agriculteurs dans chaque région : $\frac{1}{f_{i.}}$.

4.4. Construction des nuages et ajustement

Une répartition identique des cultures dans deux régions sera traduite par un même point dans un espace géométrique. Ainsi, pour mieux voir les distances entre les profils, chacun des deux tableaux se représente dans des espaces à autant de dimensions qu'il y a de modalités étudiées. On y fait figurer les point-profils dont les proximités seront révélatrices de leurs similitudes, on dit en AFC, de leurs **correspondances**, et le centre de gravité du nuage égal au profil marginal.

Tout comme dans l'ACP, un nuage sera singulier de par la dispersion des points qui le constituent. On a vu que calculer les distances entre profils était équivalent au calcul par rapport au profil moyen qui symbolise l'indépendance. La somme de ces distances est considérée comme une mesure de l'inertie du nuage de points-profils, elle-même proportionnelle au *khi2* qui est utilisée en statistique classique

pour mesurer la liaison entre deux variables qualitatives. En d'autres termes, plus le nuage aura une inertie importante, plus on s'écarte de l'indépendance.

L'AFC, va chercher un espace de représentation de plus faible dimension qui reproduit au mieux cette inertie en ajustant chacun des nuages de points-profils par des axes factoriels orthogonaux.

Réaliser l'AFC (**co**orespondence **a**nalysis) du tableau regicu :

```
> res.coa <- dudi.coa(regicu)
Select the number of axes: 2
```

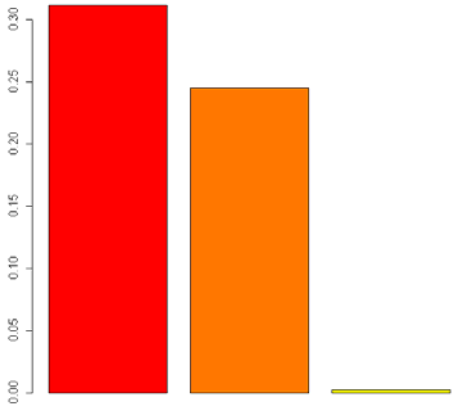


Figure 24. Diagramme des valeurs propres de l'AFC du tableau regicu.

4.5. La dualité

Les deux nuages jouent des rôles symétriques puisque ce sont les mêmes individus (les agriculteurs) qui sont représentés mais selon des optiques différentes (leur répartition au sein des régions et celle par rapport à leurs cultures). Les espaces de représentations sont similaires (on représente des fréquences qui sont positives et varient entre 0 et 1) et il est naturel de penser que l'on peut vouloir représenter les profils lignes et colonnes dans un espace commun. Cela tombe bien, car l'AFC admet cette **représentation** dite **simultanée** qui permet de visualiser les proximités entre un profil ligne par rapport à tous les profils colonnes et entre un profil colonne par rapport à tous les profils lignes.

Voyons comment sont positionnés les points profils :

De manière simple, on superpose le plan factoriel des profils lignes sur celui des profils colonnes, de telle manière que chaque région (profils lignes) se trouve au centre de gravité (**barycentre**) des positions des cultures (profils colonnes), ces dernières pondérées par leur importance dans les régions.

On fait de même avec le nuage des profils colonne où chaque culture (profils colonnes) se trouve au barycentre des positions des régions (profils lignes), ces dernières pondérées par l'importance de la culture représentée.

On obtient deux représentations dites barycentriques que l'on va superposer (à un coefficient près de mis à l'échelle des axes) pour obtenir la **représentation simultanée** (voir Figure 25)**.**

Il faudra être vigilant lors de la lecture des proximités entre points-lignes et points-colonnes. Afin de se prémunir d'éventuelles erreurs de lecture, on se méfiera des proximités entre modalités proches de l'origine du nuage en s'assurant à l'aide du cos^2 (indice de la qualité de représentation) que les points concernés sont bien représentés.

4.6. Interprétation d'une AFC

Globalement l'interprétation d'une AFC se fait de la même manière qu'une ACP en ce qui concerne les aides à l'interprétation.

```
> scatter(res.coa)
```

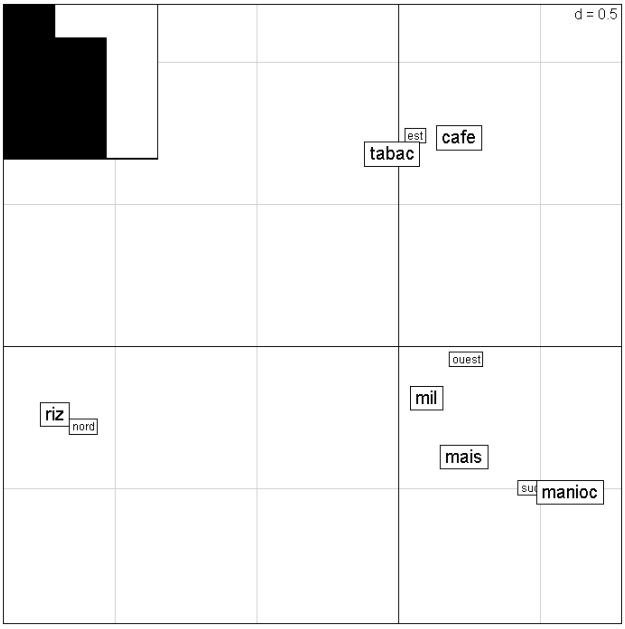


Figure 25. Représentation simultanée sur plan factoriel 1-2 des régions et des cultures.

La fonction générique score (ici `score.coa`) permet une représentation axe par axe des positions des régions et des cultures (Figure 26) :

```
> score(res.coa, xax=1, dotchar=T)
```

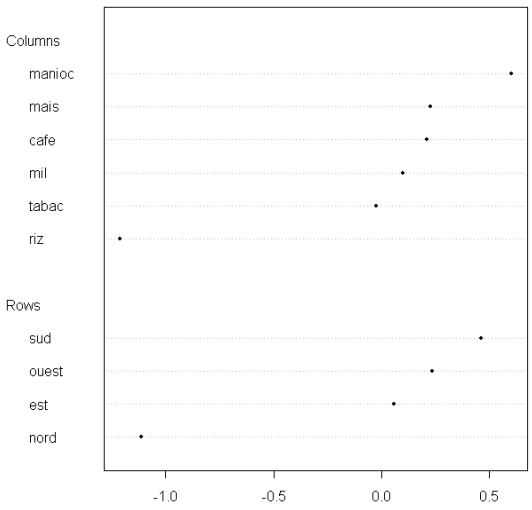


Figure 26. Représentation des cultures et des régions sur le premier facteur.

Le plan factoriel simultané des cultures et des régions (Figure 25) et les aides à l'interprétation de qualité de représentation permettent de lire les associations entre les différents types de cultures et les régions, deux modalités qui s'associent, s'attirent :

- on voit que la région Nord se distingue nettement des autres du fait de la prééminence du riz, culture qui est beaucoup plus présente dans cette région que dans les autres,
- le tabac et le café sont deux cultures qui ont la préférence des agriculteurs de l'Est,
- le Sud est une région, où l'on cultive plus facilement le manioc qui est souvent associé au mais,
- Enfin, l'Ouest est une région « moyenne » puisque elle se trouve quasiment au centre de gravité des cultures, c'est à dire qu'aucune d'entre elles n'est privilégiée dans cette région.

Nous reviendrons plus en détail, sur l'interprétation des résultats d'une analyse des correspondances dans le chapitre dédié à l'AFCM.

4.7. L'inertie

Le pourcentage d'inertie projetée sur un faible de nombre de facteurs (diagramme des valeurs propres) nous renseigne sur la forme du nuage de points en indiquant la présence d'une structure plus ou moins forte dans le tableau de données.

La Figure 27 est obtenue grâce à la commande suivante :

```
> barplot(res.coa$eig, col="grey", xlab="facteurs", ylab="valeur propre",
names.arg=c("F1","F2","F3"))
```

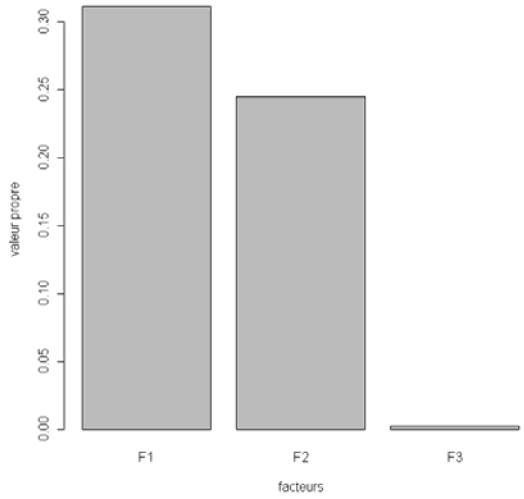


Figure 27. Diagramme des valeurs propres.

Les éléments chiffrées d'aide à l'interprétation de l'ajustement s'obtiennent à l'aide de la même fonction que celle utilisée pour l'ACP :

```
> aides.coa <- inertia.dudi(res.coa)
> aides.coa
$TOT
      inertia       cum     ratio
1 0.311180523 0.3111805 0.5573279
2 0.245077819 0.5562583 0.9962651
3 0.002085355 0.5583437 1.0000000
```

L'inertie totale (la somme des valeurs propres), quant à elle, est une mesure de liaison des deux variables. On multiplie l'inertie totale par l'effectif total de la table de contingence :

```
> obskhi <- res.coa$N*sum(res.coa$eig)
[1] 2245.1
```

Cette statistique, sous l'hypothèse d'indépendance des deux variables régions et cultures, est issue d'une distribution régit par la loi du *khi2* à 15 degrés de libertés ((p-1) × (n-1) avec p et n respectivement le nombre de lignes et de colonnes du tableau de contingence).

Calculons la valeur théorique critique au niveau de signification, disons de 95% :

```
> qchisq(0.95, 15)
[1] 24.99579
```

La probabilité qu'une estimation de la valeur `obskhi` soit plus petite que la valeur théorique critique calculée ci-dessus est :

```
> 1-pchisq(obskhi, 15)
[1] 0
```

On peut donc réfuter l'hypothèse d'indépendance sans presque aucun risque de se tromper.

5. Une méthode pour traiter les données d'enquête : l'Analyse Factorielle des Correspondances Multiples (AFCM)

L'AFCM ou ACM est une généralisation de l'AFC pour l'étude de plus de deux variables qualitatives. Les tableaux issus d'enquête par questionnaire sont souvent soumis à cette méthode qui s'avère très adaptée car elle permet de mettre en relief des systèmes de relations entre variables difficiles ou impossibles (liaisons non linéaires) à percevoir avec les méthodes descriptives classiques.

A partir d'un exemple d'enquête fictif, nous allons voir comment sont spécifiquement codées les données, quels sont les calculs effectués et surtout comment interpréter les résultats numériques et graphiques.

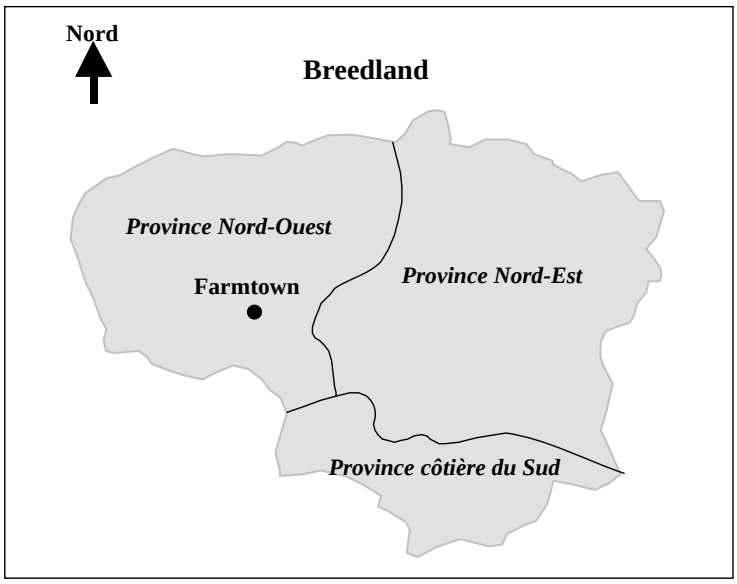


Image 2. Les provinces du Breedland.

Nous avons interrogé 50 éleveurs du Breedland répartis dans tout le pays. Ces éleveurs ont répondu à un certain nombre de questions concernant leur élevage, qu'il s'agisse de leur famille, leur troupeau, leurs pratiques d'élevages et les problèmes qu'ils rencontrent au quotidien. Nous avons recueilli toutes ces informations afin de synthétiser notre connaissance des éleveurs du Breedland au moyen d'une typologie.

Tableau 4. Dictionnaire des variables de l'enquête éleveurs du Breedland.

Les éleveurs du Breedland			
Variables actives		**modalités**	**Description**
L'éleveur			
femmes	statut familial	fem0	Célibataire
		fem1	Marié
		fem2	Polygame
enfants	Nombre d'enfants	fam0	Pas d'enfant
		fam1	entre 1 et 9 enfants
		fam2	+ de 10 enfants
deplacement	Déplacement	sed	Sédentaire
		trans	Transhumant (partiellement)
autreact	Autre activité pratiquée	agri	Agriculture
		com	Commerce
		artis	Artisanat
		aucun	Aucune
Taille et diversité du troupeau			
bovinsq	Nombre de bovins	*quantitative*	
bovins	Nombre de bovins	bov1	entre 5 et 10 bovins
		bov2	entre 11 et 50 bovins
		bov3	entre 51 et 100 bovins
		bo4	+ de 100 bovins
typebov	type d'élevage	viande	Viande
		lait	lait
		mixte	mixte
ptirum	Possession de petits ruminants	ptiruo	oui
		ptirun	non
Pratiques d'élevage			
gardien	gardiennage	gardf	Familial
		gards	Salarié
lots	mises en lots des bovins	lotso	oui
		lotsn	non
vaccination	vaccinations des animaux	Vacco	oui
		Vaccn	non
		Vaccp	parfois
vermifugation	vermifugation des animaux	Vermi1	1 fois par an
		Vermi2	plus d'1 fois par an
Variables supplémentaires			
Les problèmes du développement : 6 types de problèmes			
paturage	problème de pâturages	patuo	oui
		patun	non
eau	rareté de l'eau	eauo	oui
		eaun	non
pathologies	pathologie	patho	oui
		pathn	non
feux	feux	feuxo	oui
		feuxn	non
volinsecu	vols et insécurité	volo	oui
		voln	non
mouches	mouches	mouco	oui

		moucn	non
Variables illustratives			
region	Localisation géographique	ne no sc	Nord-Est Nord-Ouest Sud-Côtière

5.1. Les données

Le tableau de données sur lequel on travaille est issu d'un questionnaire d'enquête où chaque question est traduite en une variable dont les réponses[5] constituent les modalités de la variable. Ces modalités prennent une valeur entre 1 et *m* et constituent les valeurs possibles de la variable. Les questions de nature quantitative (*ex.* nombre de bovins) ont été, elles, recodées en classes. Les lignes de ce tableau représentent les individus qui ont participé à l'enquête.

Importation dans **R** du tableau de données préalablement sauvegardé en texte. Grâce à l'option `row.names`, on spécifie la première colonne du tableau comme libellé des identificateurs des lignes :

```
> eleveurs <- read.table("eleveurs.txt", sep=";", header=T, row.names = 1)
```

Voici un extrait de ce tableau dit 'tableau qualitatif à codage condensé' où l'on peut lire que l'individu numéro 2 habite dans la province du Nord-Est, qu'il est marié, une famille de taille moyenne, qu'il est éleveur transhumant et n'a aucune autre activité, etc.

```
> eleveurs[1:10,1:9]
     region femmes enfants deplacement autreact bovins bovinsq typebov  ptirum
1  nord-est   fem1    fam2       trans    aucun    bov3     60  viande ptrirun
2  nord-est   fem1    fam2       trans    aucun    bov3     70  viande ptrirun
3  nord-est   fem1    fam3       trans    aucun    bov3     55  viande ptrirun
4  nord-est   fem1    fam2       trans    aucun    bov4    110  viande ptriruo
5  nord-est   fem2    fam2       trans    aucun    bov4    200   mixte ptrirun
6  nord-est   fem1    fam3         sed    aucun    bov3     98  viande ptrirun
7  nord-est   fem2    fam2         sed    aucun    bov3     74  viande ptrirun
8  nord-est   fem2    fam2       trans    aucun    bov3     79   mixte ptrirun
9  nord-est   fem2    fam3       trans    aucun    bov3     82    lait ptrirun
10 nord-est   fem2    fam3       trans    aucun    bov4    140  viande ptrirun
```

La liste complète des variables :

```
> names(eleveurs)
 [1] "region"        "femmes"        "enfants"       "deplacement"
 [5] "autreact"      "bovinsq"       "bovins"        "typebov"
 [9] "ptirum"        "gardien"       "lots"          "vaccination"
[13] "vermifugation" "paturage"      "eau"           "pathologies"
[17] "feux"          "volinsecu"     "mouches"
```

5.2. Apurement et homogénéisation

Il est important d'opérer un premier classement des variables par grands thèmes d'interrogation (Tableau 4). Ne pas tout mélanger afin de rendre cohérents les résultats des analyses et faciliter par là même l'interprétation.

Surtout, on doit prendre une décision quant aux variables qui nous semblent pertinentes et les plus à même à décrire les éleveurs. Dans la pratique, c'est soit la connaissance des données[6], soit un premier essai d'analyse qui permet de le faire.

[5] ce qui suppose que les réponses aux questions soient plutôt à choix restrictifs. La reformulation des réponses d'une question ouverte en un nombre limité de modalités n'étant pas toujours très facile.

[6] Cette connaissance rentre pour une grande partie dans les choix méthodologiques et l'interprétation.

Avant de jeter nos données en pâture au logiciel d'analyse de données, il y a plusieurs préalables indispensables. Une première exploration des données va nous permettre de :

1. De visualiser les distributions des variables pour repérer celles qui présentent des données trop extrêmes ou des modalités trop peu abondantes. Celles-ci occupent alors un rôle central dans l'analyse et masquent l'essentiel de la structure du tableau de données.

2. Repérer évidemment les distributions qui ne varient pas. Les variables en question n'étant porteuses d'aucune information.

3. Faire le point sur les données manquantes. Les méthodes factorielles n'admettent pas ou très mal les non-réponses.

4. L'AFCM ne s'intéresse qu'aux variables qualitatives. Il s'agit donc de procéder à un recodage des variables continues en classes si l'on veut les garder dans le tableau de données soumis à l'analyse. Ceci dit l'AFCM autorise les variables quantitatives sous statut de variables supplémentaires mais tous les logiciels ne le permettent pas.

On sauvegarde la variable quantitative `bovinsq` dans un objet :

```
> bovinsq <- eleveurs[,"bovinsq"]
```

Les variables supplémentaires :

```
> varsup <- eleveurs[,c("paturage","eau","pathologies","feux","volinsecu",
"mouches")]
```

Ainsi que la variable illustrative `region` :

```
> region <- eleveurs$region
```

La sauvegarde des variables annexes permet de construire le data.frame des variables actives :

```
> data.actif <- eleveurs[,-c(1,6,14:19)]
> summary(data.actif)
  femmes    enfants    deplacement  autreact    bovins      typebov       ptirum
 fem0:11   fam0: 9   sed  :37    agri :17   bov1:10   lait  :11   ptirun:39
 fem1:26   fam1:32   trans:13    artis: 5   bov2:20   mixte :12   ptiruo:11
 fem2:13   fam2: 9               aucun:18   bov3:16   viande:27
                                 com  :10   bov4: 4
  gardien      lots     vaccination vermifugation
 gardf:21   lotsn:28   vaccn:11    vermi1:16
 gards:29   lotso:22   vacco:37    vermi2:34
                       vaccp: 2
```

Avant de passer à l'interprétation des résultats, nous allons rapidement voir comme pour les autres méthodes, quels sont les calculs qui sont opérés.

5.3. Transformation du tableau de données

L'AFCM est une généralisation de l'AFC puisqu'elle permet d'étudier les relations entre plus de deux variables qualitatives. Pour ce faire, on opère une transformation du tableau initial à codage condensé en **tableau disjonctif complet**. Celui-ci présente les valeurs des variables sous la forme 0 et 1, selon que l'individu possède ou non la modalité considérée :

```
> disj <- acm.disjonctif(data.actif)
```

Un extrait du tableau disjonctif complet :

```
> disj[1:5,7:12]
   deplacement.sed deplacement.trans autreact.agri autreact.artis
5                0                 1             0              0
6                1                 0             0              0
7                1                 0             0              0
8                0                 1             0              0
```

```
9                0                1             0             0
10               0                1             0             0
   autreact.aucun autreact.com
5                1            0
6                1            0
7                1            0
8                1            0
9                1            0
10               1            0
```

Sous cette forme, on peut traiter un nombre quelconque de variables qualitatives ou quantitatives (ces dernières ayant été recodées en qualitatives préalablement). Le fait d'appréhender des données quantitatives en classes de valeurs (modalités) est forcément synonyme d'information plus pauvre ou grossière mais elle a l'avantage premièrement,

- de faciliter l'interprétation (par exemple les modalités de la variable taille de troupeau de bovins : petit, moyen, grand sont d'un usage plus fréquent en terme de description),

- deuxièmement d'appréhender les éventuelles liaisons non linéaires entre variables (ce qui n'est pas le cas de l'ACP qui s'appuie sur les corrélations linéaires entre variables).

De plus, ce codage permet de conserver l'entité individu (qui n'existe pas en AFC) et de pouvoir visualiser la répartition des individus sur leur propre plan factoriel. Ce qui constitue une première étape à l'élaboration des typologies.

Réalisons le produit matriciel du tableau `disj` par son transposé :

```
> burt <- t(as.matrix(disj))%*%(as.matrix(disj))
```

Tableau obtenu également avec la fonction suivante de la librarie ade4 :

```
> burt <- acm.burt(data.actif, data.actif)

> burt[7:12,7:12]

                   deplacement.sed deplacement.trans autreact.agri
  deplacement.sed               37                 0            17
  deplacement.trans              0                13             0
  autreact.agri                 17                 0            17
  autreact.artis                 3                 2             0
  autreact.aucun                 7                11             0
  autreact.com                  10                 0             0

                   autreact.artis autreact.aucun autreact.com
  deplacement.sed               3              7           10
  deplacement.trans             2             11            0
  autreact.agri                 0              0            0
  autreact.artis                5              0            0
  autreact.aucun                0             18            0
  autreact.com                  0              0           10
```

Le résultat est un tableau qui juxtapose toutes les tables de contingence, appelé tableau de Burt. Finalement, l'AFCM s'apparente à l'analyse des correspondances de cette « super-table » de contingence.

5.4. Construction des nuages et ajustement

Le principe de l'AFCM tient en une phrase : AFC sur tableau disjonctif complet. Ainsi, à partir de tableau de données logiques, on va définir des profils-lignes (individus) et des profils-colonnes (modalités) qui vont se matérialiser chacun en un nuage de points représentable dans des espaces de dimension élevée.

De la même manière que pour l'AFC, chacun de ces deux nuages dont les formes sont révélatrices de leur information, vont être résumé par des facteurs. Ce qui constitue l'essence d'un nuage de points, ce sont les distances entre les points et

c'est celle du *khi2* qui est définie pour évaluer la proximité entre deux individus ou entre deux modalités.

On définit donc une distance entre deux **points-individus** *i* et *i'* :

$$d^2(i,i') = \frac{1}{s}\sum_{j=1}^{p}\frac{n}{z_{.j}}(z_{ij} - z_{i'j})^2$$

Équation 4. Distance entre 2 points individus issus d'un tableau disjonctif complet.

et la distance entre deux **points-modalités** *j* et *j'* :

$$d^2(j,j') = \sum_{i=1}^{n} n\left(\frac{z_{ij}}{z_{.j}} - \frac{z_{i'j}}{z_{.j''}}\right)^2$$

Équation 5. Distance entre 2 points modalités issus d'un tableau disjonctif complet.

avec,

- $z_{ij} = 1$ ou 0 selon que l'individu possède la modalité *j* de la question *q*,
- $z_{.j}$ la somme marginale colonne $\left(z_{.j} = \sum_{i=1}^{n} z_{ij}\right)$: nombre d'individus ayant choisi la modalité j de la question q.
- *s* le nombre de questions-variables du tableau à codage condensé $\left(s = z_{i.} = \sum_{j=1}^{p} z_{ij}\right)$.
- *n* le nombre d'individus total .

Ainsi, deux points-individus sont proches géométriquement s'ils ont choisi les mêmes modalités. Ils sont éloignés, s'ils n'ont pas répondu de la même manière. On peut remarquer, qu'une modalité est d'autant plus importante dans le calcul de cette distance que son abondance est faible (poids important $z_{.j}$ des modalités rares).

Les deux nuages peuvent faire l'objet d'une représentation simultanée même si en pratique on le fait rarement, on peut toutefois retenir qu'une modalité est positionnée sur un plan factoriel au barycentre des scores des individus qui la possèdent.

5.5. L'interprétation des résultats

L'AFCM se fonde dans son principe sur l'AFC mais des règles spécifiques régissent l'interprétation des résultats. L'analyse du tableau de données des éleveurs du Breedland, va nous permettre de dépouiller les résultats numériques et graphiques d'une AFCM.

Lancer l'ACM et choisir le nombre de facteurs après visualisation du diagramme des valeurs propres :

```
> res.acm <- dudi.acm(data.actif)
Select the number of axes: 3
```

5.5.1. Le diagramme des valeurs propres

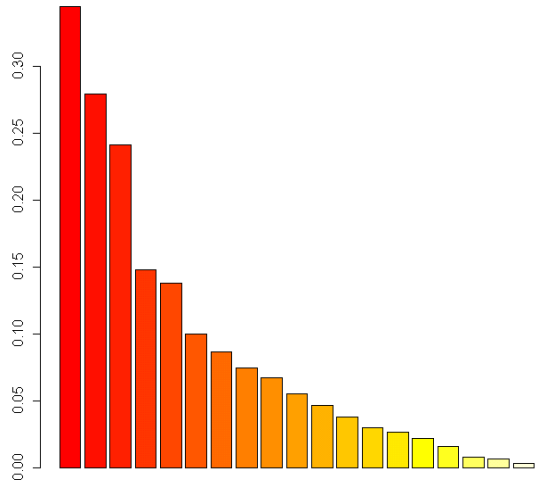


Figure 28. Diagramme des valeurs propres de l'ACM du tableau eleveurs.

C'est l'indicateur privilégié pour visualiser la part d'information (en l'occurrence, les liaisons) qui a été résumée par chacun des facteurs. Mais en AFCM, l'inertie totale ne dépend pas des liaisons mais du nombre de modalités. Il ne faut donc pas s'y référer mais plutôt observer **le profil de décroissance** du diagramme qui indique dans quelle mesure le tableau de données est structuré.

Ici, on observe un saut après le troisième bâton, les facteurs suivants sont d'un intérêt secondaire et l'on retient donc 3 axes significatifs. Les facteurs principaux sont souvent des gradients qui ordonnent les individus selon une ou plusieurs variables prépondérantes, ou bien des facteurs qui partagent les individus en groupes et mettent en évidence leurs oppositions.

L'effet taille ou effet Guttman

Comme pour l'ACP, les variables qualitatives peuvent aller dans le même sens, c'est à dire être toutes très associées et donc se réduire à une seule dimension. Cela arrive très souvent, lorsque les modalités des variables peuvent être ordonnées (par exemple les variables originellement quantitatives que l'on a mis en classes).

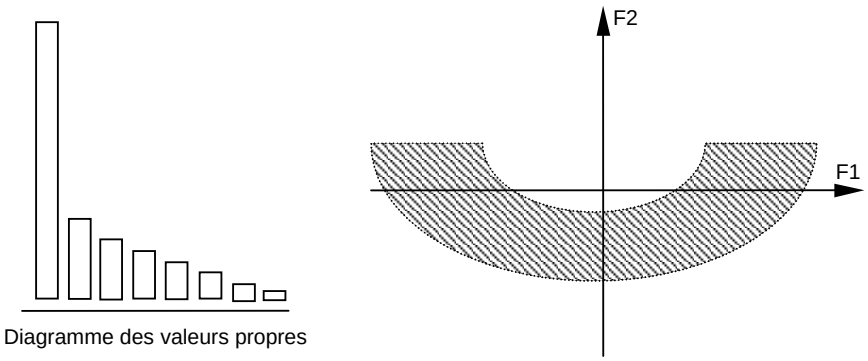


Figure 29. L'effet Guttman en AFC ou AFCM.

On peut détecter cet effet taille grâce au diagramme des valeurs propres qui fait ressortir le premier bâton de manière significative et sur le premier plan factoriel (facteur 1/facteur 2) un positionnement des modalités et des individus selon une parabole (Figure 29).

L'interprétation d'un tel plan factoriel se résume à un axe prépondérant (le premier, F1) qui est un facteur d'échelle (gradient qui ordonne les individus de ceux qui prennent des petites valeurs à ceux qui prennent de grandes valeurs pour les variables les plus contributives) et un second (F2) qui distingue les situations extrêmes aux situations moyennes.

5.5.2. Les aides numériques à l'interprétation

Les variables et leurs modalités actives

L'AFCM est une analyse factorielle, elle cherche à rendre compte de ce qui distingue principalement les éleveurs. Les variables entrent pour une part plus ou moins grande dans cette distinction, les **contributions ou contributions absolues** permettent d'en juger et d'interpréter les facteurs.

```
> inertia.dudi(res.acm, row.inertia=T, col.inertia=T)$col.abs
                     Comp1 Comp2 Comp3
femmes.fem0            939    17    39
femmes.fem1             30   169    20
femmes.fem2            419   212   146
enfants.fam0           902    41    81
enfants.fam1            35     0    48
enfants.fam2           358    43    16
deplacement.sed        292    52    98
deplacement.trans      830   148   278
autreact.agri           52    79  1646
autreact.artis         220   469   976
autreact.aucun         602   124   213
autreact.com           169  1752   126
bovins.bov1            474   604   582
bovins.bov2            338   298   246
bovins.bov3            795     2   118
bovins.bov4            367    10   348
typebov.lait             9  1472   123
typebov.mixte          207   228  1080
typebov.viande         133   208   840
ptirum.ptrirun          41   204   362
ptirum.ptriruo         145   723  1283
gardien.gardf          487   343   265
gardien.gards          353   248   192
lots.lotsn             305   202   145
lots.lotso             388   257   184
vaccination.vaccn      677   570    21
vaccination.vacco      105   222     1
vaccination.vaccp      288    66    42
vermifugation.vermi1    29   840   330
vermifugation.vermi2    14   395   155
```

Dans la pratique, après avoir détecté les variables qui pèsent le plus dans l'explication des facteurs, on s'intéresse surtout aux modalités. Sans qu'il n'existe véritablement de règles de dépouillement, on peut choisir de ne retenir que les modalités ayant une contribution supérieure à la moyenne (sachant que la contribution totale est de 10000), soit dans notre cas où l'on dispose de 30 modalités :

```
> 10000/30
[1] 333.3333
```

En classant par ordre décroissant, on lit que les plus fortes contributions sur le premier facteur sont,

```
> sort(inertia.dudi(res.acm, row.inertia=T, col.inertia=T)$col.abs[,1],
decreasing=T)
```

```
          femmes.fem0          enfants.fam0     deplacement.trans
                  939                   902                   830
          bovins.bov3     vaccination.vaccn        autreact.aucun
                  795                   677                   602
       gardien.gardf           bovins.bov1          femmes.fem2
                  487                   474                   419
          lots.lotso           bovins.bov4         enfants.fam2
                  388                   367                   358
       gardien.gards           bovins.bov2           lots.lotsn
                  353                   338                   305
     deplacement.sed     vaccination.vaccp        autreact.artis
                  292                   288                   220
       typebov.mixte          autreact.com       ptirum.ptriruo
                  207                   169                   145
      typebov.viande     vaccination.vacco         autreact.agri
                  133                   105                    52
       ptirum.ptirun          enfants.fam1          femmes.fem1
                   41                    35                    30
vermifugation.vermi1 vermifugation.vermi2         typebov.lait
                   29                    14                     9
```

En général, lorsque la contribution d'une modalité est forte pour un axe, cette dernière est bien représentée. On s'en assure en lisant les **cos²**, indicateur de la qualité de leur représentation sur l'axe concerné (elles sont additives par axes, on peut donc en déduire la qualité sur un plan).

De manière générale, tous ces résultats ne sont pas examinés dans le détail et l'on se contente de réaliser les synthèses à l'aide des représentations graphiques factorielles qui sont dans bien des cas suffisamment parlantes.

Les individus actifs

De la même manière que pour les variables, certains éleveurs parce qu'ils sont un petit nombre à avoir répondu de manière commune à certaines questions, ont une position plus excentrée sur le plan factoriel. Ils ont ainsi « tiré » l'axe vers eux et contribué à la détermination du facteur.

```
> sort(inertia.dudi(res.acm, row.inertia=T, col.inertia=T)$row.abs[,1],
decreasing=T)
 18  16  10   1   4   9   3  25  27  41  14   5   2   8  11  36  23  44  47
972 615 573 570 519 486 453 423 423 396 374 360 346 313 309 279 267 244 202
 12  30  15  17  34   7  39  40  20  42  13  19   6  31  48  32  24  26  29
177 175 154 154 139 134 108  81  80  80  78  69  58  49  40  38  33  33  33
 33  35  37  21  45  49  50  43  28  38  46  22
 28  28  28  22  22  16  10   5   3   1   1   0
```

Sur le facteur 1, on recense les éleveurs à forte contribution : 18, 16, 10, 1, 4, 9, 3, 25, 27. Ces éleveurs qui ont une contribution importante à l'inertie projetée sont ceux qui s'écartent le plus de « l'éleveur moyen ».

5.5.3. Les représentations graphiques

On a déterminé le nombre d'axes permettant de visualiser graphiquement les deux nuages de points. On peut afficher les plans F1-F2, F2-F3 et F1-F3 chacun pour le nuage des modalités et des individus en utilisant pour leur lecture les aides à l'interprétation numérique[7].

Le plan factoriel des individus (Figure 30) est intéressant pour détecter d'éventuels points extrêmes et/ou aberrants (*1*), une partition des individus (*2*), ou une répartition homogène (*3*).

[7] On peut bien sûr comme pour l'AFC faire une représentation simultanée des individus et des modalités, mais en AFCM elle est de peu d'intérêt sauf si les individus ne sont pas « anonymes ».

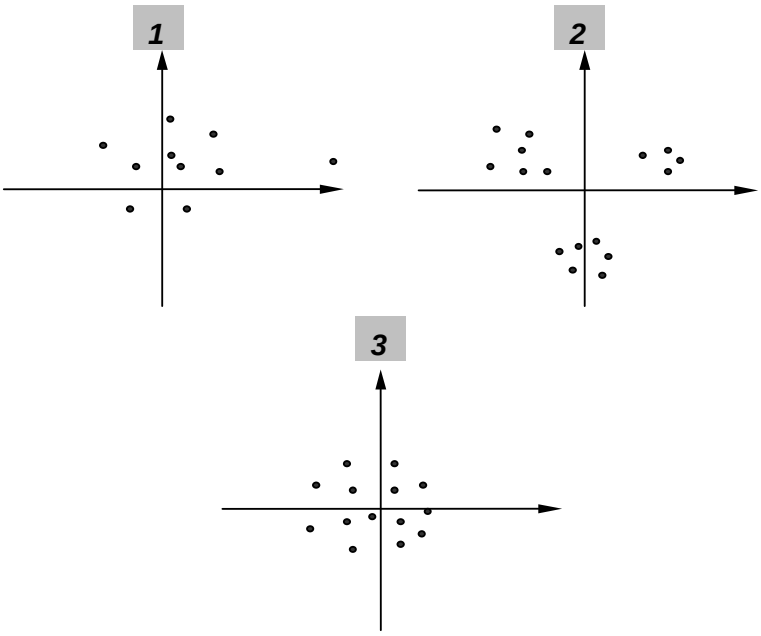


Figure 30. Différentes configurations de la répartition des individus sur les plans factoriels.

En ce qui concerne, le plan factoriel des modalités, on s'intéresse surtout au plan F1-F2 révélateur des proximités principales. On regardera éventuellement sur les plans suivants, pour voir s'il n'existe pas des associations « secondaires » qui peuvent toutefois être intéressantes.

On représente successivement le plan factoriel 1-2 des individus :

```
> s.label(res.acm$li, xax=1, yax=2, clabel=0.6)
```

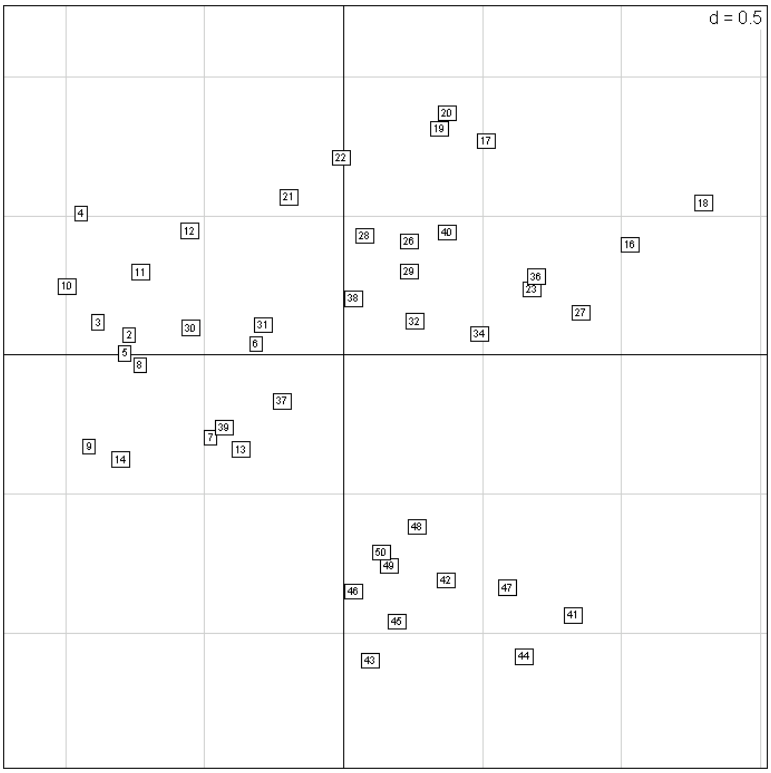


Figure 31. Plan factoriel 1-2 des individus

On voit nettement la juxtaposition de plusieurs amas de points montrant bien qu'il existe une partition des individus. Les deux premiers axes séparent ces groupes, il reste à expliquer les axes factoriels à l'aide des modalités les plus influentes afin de comprendre ce qui distingue les groupes d'éleveurs.

Puis celui des modalités :

```
> s.label(res.acm$co, xax=1, yax=2, clabel=0.6)
```

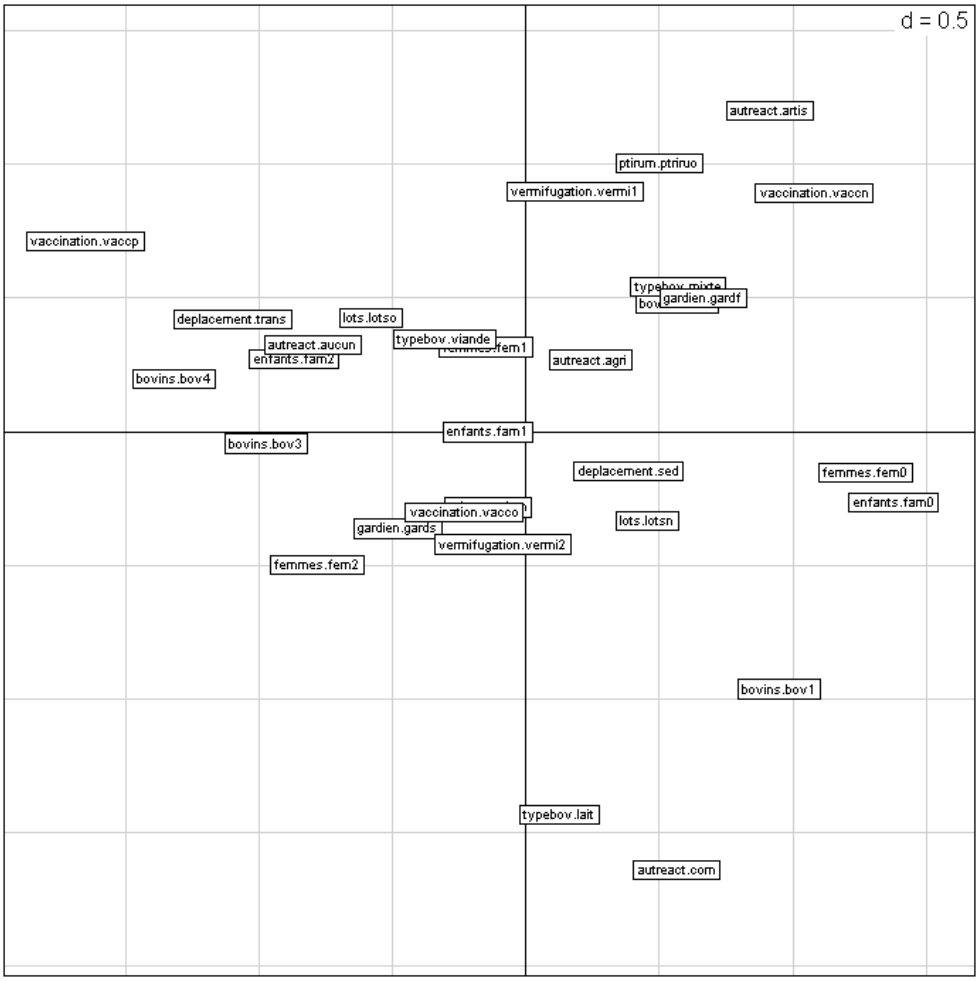


Figure 32. Le plan factoriel 1-2 des modalités.

Des modalités s'associent lorsque les mêmes éleveurs ont répondu de façon identique aux mêmes questions. Pour la lecture de ce plan factoriel, on s'intéresse surtout à la périphérie du nuage. Les modalités proches de l'origine n'apportent que très peu d'informations quant à la distinction des éleveurs. Si certaines d'entre elles en apportent, il faudra visualiser leurs positions sur d'autres axes où leurs contributions sont importantes.

Les modalités les plus contributives (voir plus haut) se trouvent aux extrémités des axes, et celles comme bov1 qui se trouvent sur une bissectrice ont des contributions relativement importantes sur les deux axes.

Des utilitaires graphiques permettent de dresser une synthèse graphique du poids des modalités dans la constitution d'un facteur. La Figure 33 montre que le score d'une modalité est en quelque sorte une moyenne de la position des individus la possédant. Utilisation avec le facteur 3 :

```
> score.acm(res.acm, 3)
```

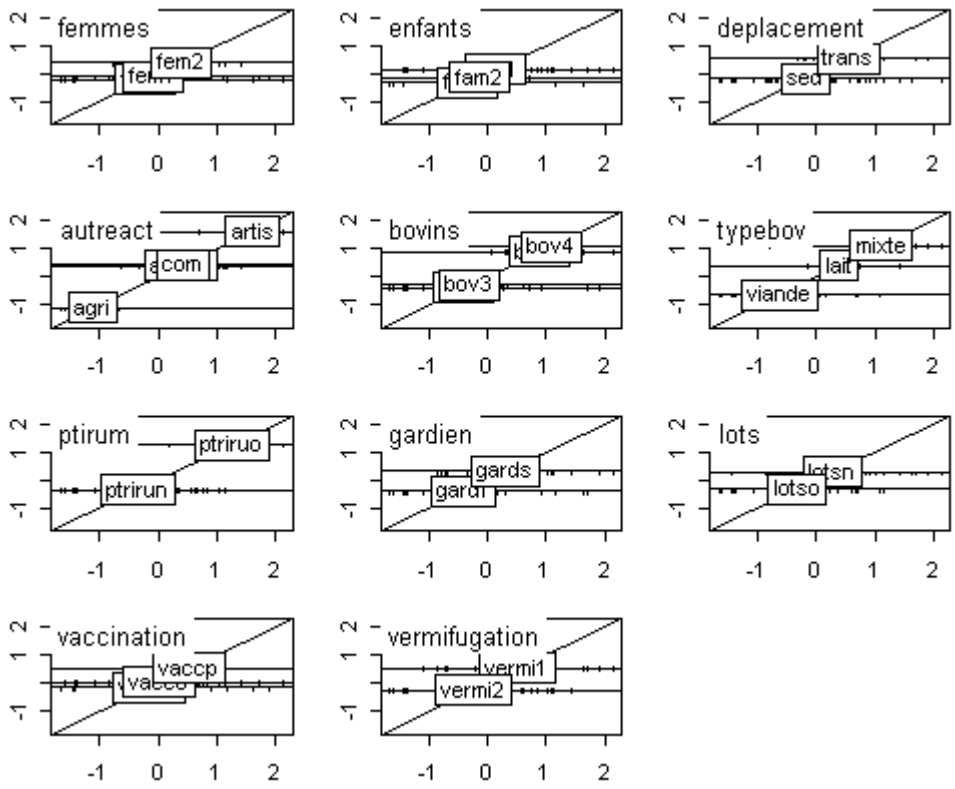


Figure 33. Synthèse des scores des modalités sur le facteur 3 à l'aide de la fonction graphique score.acm.

Sur la Figure 33, le facteur 3 est positionné en abscisse et en ordonnée de chaque sous graphique. On represente ainsi pour chaque modalité, la variabilité (et la moyenne, point-modalité) des scores des éleveurs qui la possède. Les points-modalités proches de zéro sont peu contributifs au facteur, au contraire de ceux qui se positionne de part et d'autre de l'axe.

Cette synthèse graphique par projection des moyennes des scores des individus symbolisés par les points modalités est également disponible pour le plan F1-F2 grâce à la fonction :

```
> scatter.acm(res.acm)
```

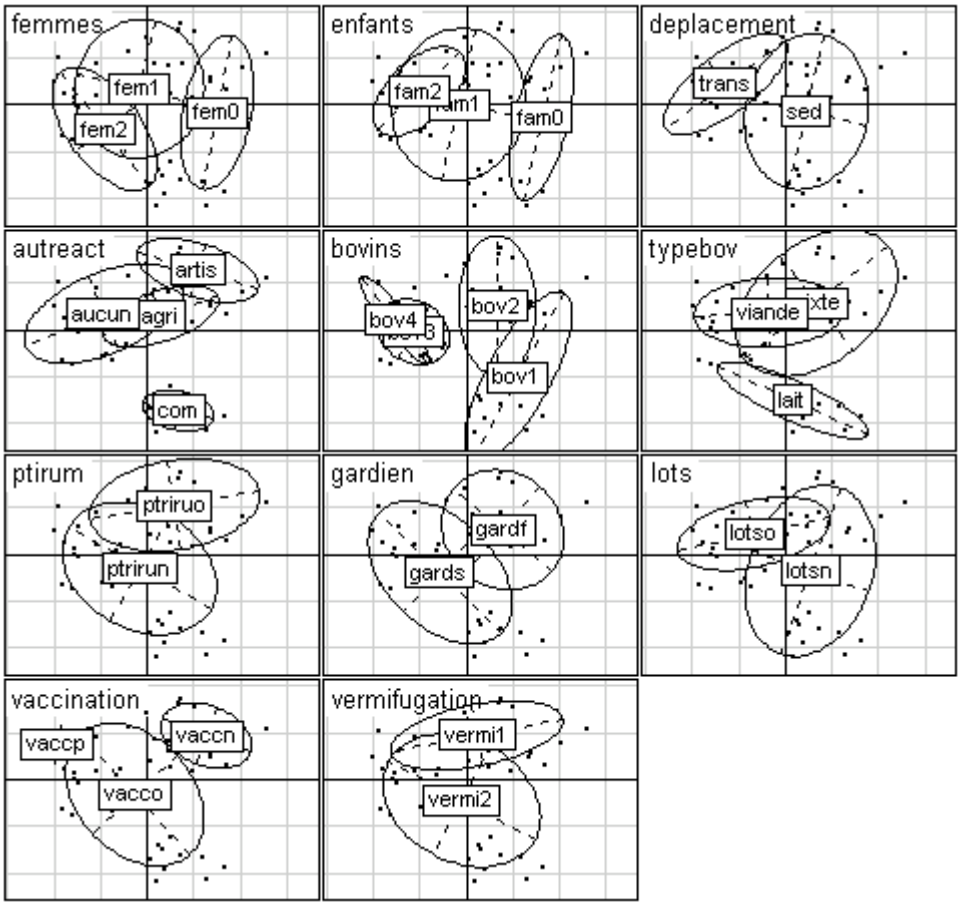


Figure 34. Synthèse des scores des modalités sur le plan factoriel 1-3 à l'aide de la fonction graphique scatter.acm.

Sur la Figure 34, le plan factoriel des individus est reproduit pour chaque variable active. On représente dans chaque sous-fenêtre la moyenne et la variabilité des scores (à l'aide d'une ellipse d'inertie) des éleveurs possédant chaque modalité.

```
> s.label(res.acm$co[c(9:12,13:16),], xax =1, yax=2, clabel=0.8)
```

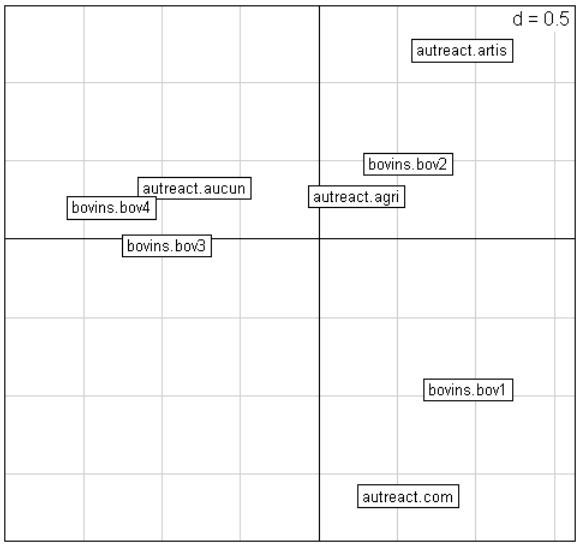


Figure 35. Deux variables qualitatives associées sur le plan factoriel 1-2.

On peut relier les modalités des variables (ex. `bovins` et `autreact`). La liaison (linéaire ou pas) entre deux variables existe lorsque les droites reliant les modalités sont parallèles (Figure 35).

Grâce aux aides à l'interprétation numériques et au plan factoriel, voici ce que l'on peut dire de cette analyse.

5.5.4. Description des facteurs

Elle est réalisée à l'aide des représentations graphique, notamment les différents plans factoriels des modalités retenus pour l'interprétation. Il conviendra dans le même temps de vérifier les contributions des modalités identifiés car leur position peut, dans certains cas, être déformée par l'opération de projection.

Premier axe :

On trouve côté négatif de l'axe F1, les modalités `deplacement.trans`, `bovins.bov3`, `bovins.bov4`, `autreac.aucun` qui sont les modalités qui correspondent aux éleveurs transhumants uniquement éleveurs possédant des troupeaux importants. Ajoutons que les familles concernées sont nombreuses (`enfants.fam2`).
Coté positif, les éleveurs célibataires (`femmes.fem0`) sans enfants (`enfants.fam0`) qui ont la particularité de ne pas vacciner leurs animaux (`vaccination.vaccn`). La variable la plus contributive (somme des contributions des modalités) sur le facteur F1 est `bovins`, la taille du troupeau est ainsi la caractéristique qui oppose principalement les éleveurs. Les deux groupes qui sont mis en évidence correspondent à des situations opposées du point de vue de la taille du troupeau.

Second axe :

En haut de cet axe F2, on trouve des éleveurs-artisan (`autreact.artis`) qui ont la particularité de vermifuger uniquement une fois dans l'année (`vermifugation.vermi1`) et de ne pas pratiquer la vaccination (`vaccination.vaccn`). Ils possèdent en sus de leurs bovins des petits ruminants (`ptirum.ptiruo`).

En bas, on trouve des éleveurs qui sont également commerçants (`autreact.com`). Les troupeaux comportent peu de têtes (`bovins.bov1`) se composant pour l'essentiel de vaches laitières (`typebov.lait`).

Ce second axe oppose deux groupes d'éleveurs mais n'est pas l'expression d'un gradient particulier. La variable la plus contributive[8] est `autreact.`

On remarque que la modalité `vaccination.vaccn` (pas de pratique systématique de la vaccination) contribue également aux axes F1 et F2, cela signifie que cette caractéristique est commune aux deux groupes d'éleveurs qui se distinguent sur ces axes.

Troisième axe :

L'interprétation du facteur 3 est réalisée à l'aide de la Figure 33.

[8] il faut se méfier car en AFCM, la contribution d'une variable est également fonction du nombre de modalités qu'elle possède.

Du côté négatif se détachent très nettement les modalités `autreact.agri` et `typebov.viande`. Ainsi, on distingue un groupe d'agro-éleveurs dont les bêtes sont presque exclusivement élevées pour leur viande.

De la même manière, du côté négatif, on discerne nettement un groupe d'éleveurs qui pratique l'artisanat en seconde activité (`autreact.artis`), possédant des petits troupeaux (`bovins.bov1`) mais également de beaucoup plus important (`bovins.bov4`) et dont les bovins sont élevés pour la viande et le lait (`typebov.mixte`). Ils déclarent également posséder des petits ruminants (`ptirum.ptiruo`).

Attention à l'erreur de perspective induite par la représentation en deux dimensions qui rapproche `bovins.bov4` de `typebov.mixte` et `ptirum.ptiruo`. Cette association existe `bovins.bov4` mais n'est pas très bien représentée sur l'axe F3, il faut donc lire ses éventuelles autres associations sur un autre plan.

5.5.5. Les éléments supplémentaires

Un certain nombre de variables qui expriment l'opinion des éleveurs quant aux problèmes de développement ont été défini en supplémentaires.

```
> names(varsup)
[1] "paturage"    "eau"         "pathologies" "feux"
[5] "volinsecu"   "mouches"
```

Elles permettent d'aider à interpréter les facteurs sans pour autant participer à leur construction (leur contribution aux axes est nulle). Cette distinction entre éléments actifs et supplémentaires est importante. Il faut nécessairement dissocier les variables descriptives actives afin que la synthèse de leurs liaisons puissent être réalisée efficacement. Ce sont en effet les variables qui définissent les distances entre individus, il faut donc pouvoir mélanger des variables qui font conserver un sens aux distances calculées. La mise en supplémentaire permet de former des groupes de variables pour conserver une ensemble de variables actives homogènes.

Afin de calculer les coordonnées factorielles des modalités supplémentaires, il s'agira de d'effectuer soi-même la transformation préalable du tableau des variables supplémentaire. Cette transformation dépend du type d'analyse factorielle initial. Dans le cas d'une ACM, utiliser la fonction `acm.disjonctif` pour obtenir le tableau disjonctif complet des variables supplémentaires. Pour les autres analyses, consulter la documentation `help(supcol)`.

```
> res.sup <- supcol(res.acm, acm.disjonctif(varsup))
> res.sup
                        Comp1       Comp2       Comp3
paturage.patun     -0.10307437  0.24254441  0.25151395
paturage.patuo      0.10307437 -0.24254441 -0.25151395
eau.eaun            0.04739811  0.06873558 -0.22211528
eau.eauo           -0.04739811 -0.06873558  0.22211528
pathologies.patho   0.10754215  0.20022237  0.15592794
pathologies.pathon -0.10754215 -0.20022237 -0.15592794
feux.feuxn          0.25118833 -0.11542931  0.14961990
feux.feuxo         -0.25118833  0.11542931 -0.14961990
volinsecu.voln      0.02317676  0.09364824 -0.01369814
volinsecu.volo     -0.02317676 -0.09364824  0.01369814
mouches.moucn       0.17614654 -0.07816142 -0.19170543
mouches.mouco      -0.17614654  0.07816142  0.19170543
```

La Figure 36 montre qu'aucune modalité supplémentaire ne se détache du centre du plan factoriel. Leurs poids dans la caractérisation des facteurs semblent faible :

```
> s.label(res.acm$co, xax =1, yax=2, clabel=0.6)
> s.label(res.sup, 1, 2, clabel=0.7, add.plot=T)
```

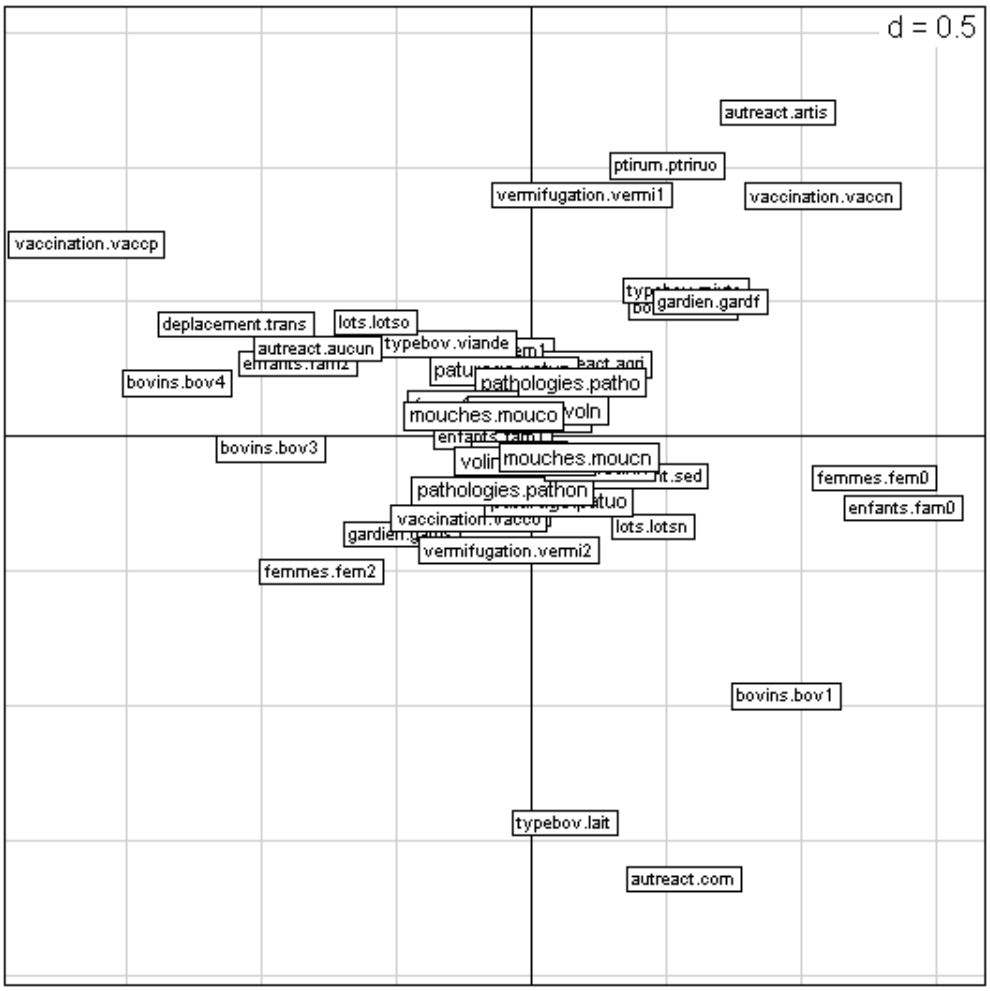


Figure 36. Plan factoriel 1-2 des modalités actives et supplémentaires (taille des caractères plus grande).

La Figure 36 n'est pas très claire du fait de la chevauchement des étiquettes des modalités. Une autre possibilité consiste à projeter les centres de gravités définis par les modalités de chaque variable supplémentaires pour chaque axe factoriel. Ce mode de représentation permet d'étudier les liens entre chaque facteur - et donc les modalités qui ont permis leur construction - et les variables supplémentaires (Figure 37) :

```
par(mfrow = n2mfrow(ncol(varsup)))
for (i in 1:(ncol(varsup)))
    s.class(res.acm$li, varsup[,i], clab = 1.5, sub = names(varsup)[i],
    csub = 2, possub = "topleft", cgrid = 0, csta = 0, cpoi = 0)

# n2mfrow : multi-fenêtrage en fonction du nb de var. supplémentaires
```

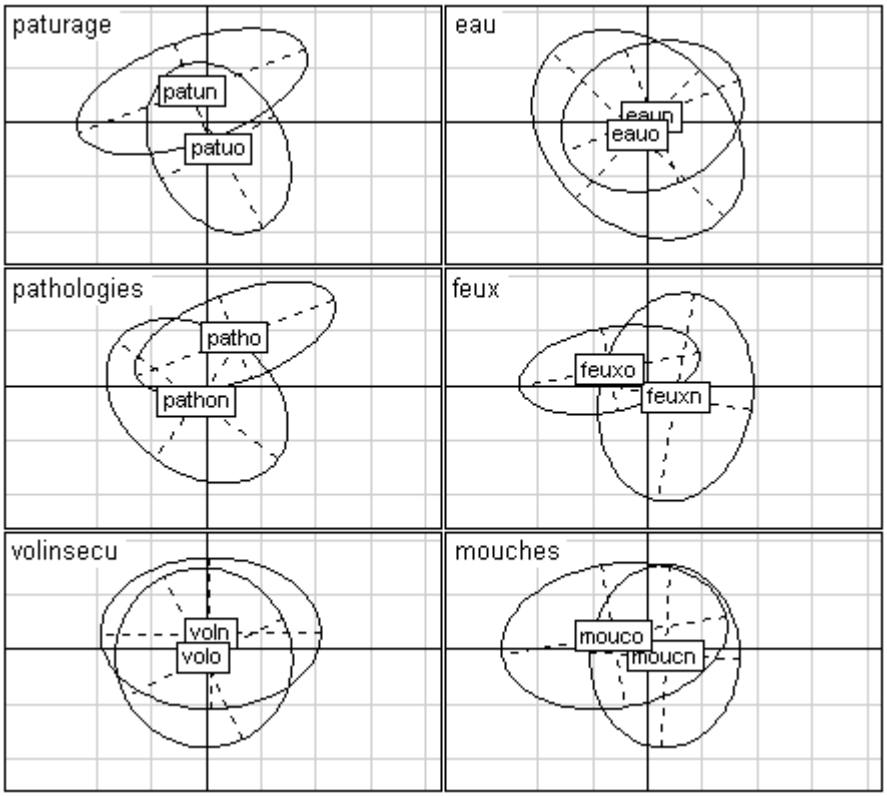


Figure 37. Représentation des liens entre variables supplémentaires et les facteurs 1 et 2 de l'ACM des éleveurs

Aucuns liens marquants entre les variables supplémentaires et les facteurs n'est à signaler même si certaines variables comme `paturage`, `pathologies` et `feux` semblent indiquer de légères préférences.

6. Les méthodes automatiques de classification

6.1. Principe

Après avoir tiré d'un tableau de données l'information principale qui caractérise les individus, on sait désormais quels sont les traits essentiels qui font que les individus se différencient. Cette connaissance acquise, on voudrait une fois de plus simplifier notre tableau afin de se référer non plus à des individus mais à des classes d'individus plus faciles à manipuler et à décrire. Cette démarche de recherche de types, de systèmes, fait partie intégrante de l'analyse des données et se concrétise en un certain nombre de méthodes dédiées à ces objectifs : les méthodes automatiques de classification.

A la grande différence des méthodes factorielles, elles font appel à une démarche algorithmique, c'est à dire une suite d'opérations qui se réitèrent jusqu'à l'obtention du résultat recherché. Donc pas de formulation mathématique complexe, juste un enchaînement simple et intuitif que nous allons décrire à partir de l'exemple des éleveurs du Breedland.

Il existe de nombreux algorithmes de classification, on peut toutefois les décomposer en deux grandes familles :

- les algorithmes qui fournissent des hiérarchies de partitions (méthodes hiérarchiques),

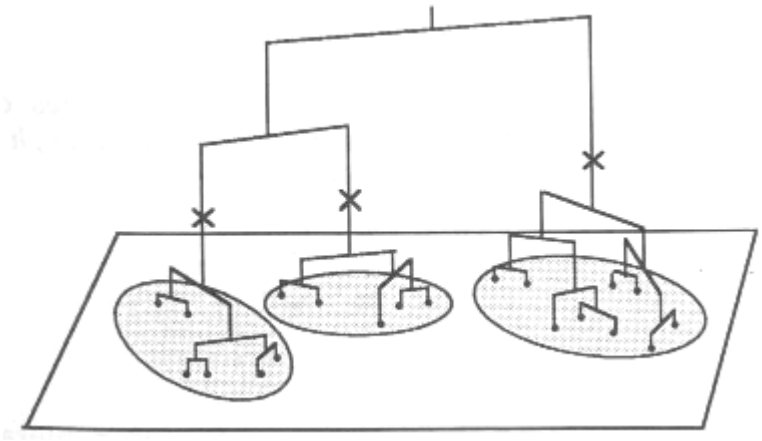


Figure 38. Représentation d'une hiérarchie de partitions.

- les algorithmes qui fournissent une seule partition optimale au sens d'un critère et dont le nombre de classes a été choisi au départ (les méthodes des nuées dynamiques).

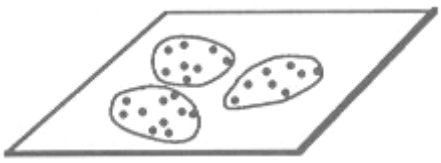


Figure 39. Partition finale obtenue par méthode nuées dynamiques.

Nous nous intéresserons uniquement à la première famille et notamment à la Classification Ascendante Hiérarchique (CAH) disponible dans la librairie `mva` du logiciel **R**.

```
> library(mva)
> library(help='mva')
```

6.2. La Classification Ascendante Hiérarchique (CAH)

Son principe est le suivant : on groupe un à un les individus ou les groupes d'individus déjà classés, selon leur ressemblance pour obtenir des partitions successives[9] et que l'on représente graphiquement sous forme d'un arbre appelé **dendrogramme ou arbre hiérarchique**. On visualise ainsi la manière dont les individus se sont groupés et on peut se faire une idée de la pertinence de telle ou telle partition.

6.2.1. Quel tableau soumettre à l'analyse ?

Les tableaux d'entrée d'une CAH peuvent être qualitatif ou quantitatif, il conviendra de choisir une mesure de distance adaptée entre deux individus. Une multitude de distances sont disponibles[10], le choix doit être fait en fonction de la manière dont on définit pour notre étude une ressemblance entre deux individus.

L'algorithme de classification hiérarchique est réalisé non pas sur le tableau de données initial mais sur la matrice de distance des individus. Cette matrice est symétrique, ses valeurs inscrites de part et d'autre de la diagonale sont identiques et la diagonale est elle-même nulle. Cette matrice se calcule à l'aide de la fonction `dist(x, method = "euclidean")` de la librairie `mva`, avec :

`x` : le tableau de données éventuellement transformées si les variables ne sont pas exprimées dans les mêmes unités,

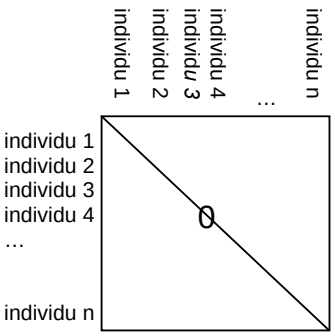
`method` : le choix d'une méthode de calcul de distance entre deux individus. Sont en autres disponibles les distances : "`euclidean`" pour un tableau de données continues et "`binary`" ou la distance de Jaccard pour les tableaux binaires lorsque la possession conjointe d'un caractère doit être valorisée.

[9] on obtient autant de partitions qu'il y a d'individus moins 1.

[10] La librarie ade4 dispose d'un ensemble très complet de fonctions et d'options pour le calcul des matrices de distances susceptible d'être ré-utilisées avec les fonctions de la librairie mva.

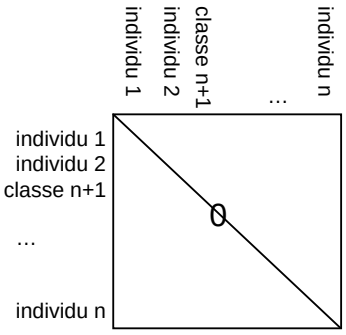
6.2.2. Le principe de l'algorithme de regroupement

1. Calcul du tableau de distance : recensement de l'ensemble des valeurs de distances entre tous les couples d'individus.



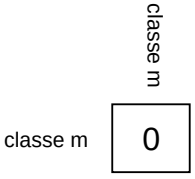
étape 1.

2. On regroupe le couple d'individus qui possède la distance la plus petite, par exemple 3 et 4, ils forment désormais la classe que l'on numérote n+1. On recalcule le tableau de distances avec une ligne et une colonne en moins car le couple précédemment réuni a disparu au profit de l'entité classe.



étape 2.

3. On réitère la procédure en agrégeant les individus ou classes qui possèdent la plus petite distance et en recalculant les tableaux de distances successifs jusqu'à l'obtention d'une seule classe qui regroupe l'ensemble des classes, c'est à dire des individus.



fin de la procédure.

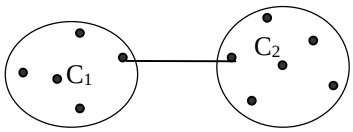
6.2.3. Choix du critère d'agrégation

La procédure de classification hiérarchique nécessite le choix d'un critère d'agrégation qui aura une utilité très concrète : celle de définir la méthode de calcul

de la distance entre un individu et une classe ou entre deux classes. Il existe plusieurs méthodes dont le choix par l'utilisateur peut avoir une influence contrastée sur la partition finale obtenue. Nous présentons ci-dessous les principaux critères d'agrégation disponible dans la fonction `hclust` de la librairie `mva`.

- Le critère du saut **minimum** : `hclust(d, method = "single")`

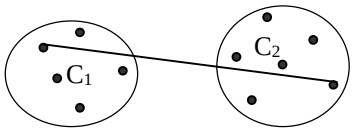
  La distance entre deux classes (ou entre une classe et un individu) est la plus petite distance qu'il peut exister entre les individus de chaque classe.



le saut minimum

- Le critère du saut **maximum** : `hclust(d, method = "complete")`

  La distance entre deux classes (ou entre une classe et un individu) est la plus grande distance qu'il peut exister entre les individus de chaque classe.



le saut maximum

- Le critère de la **moyenne** : `hclust(d, method = "average")`

  La distance entre deux classes (ou entre une classe et un individu) est la moyenne des distances entre les individus de chaque classe.

- Le critère du saut de **Ward** : `hclust(d, method = "ward")`

  On s'attarde sur ce critère qui est très utilisée en CAH car il est fondé sur l'idée même de la stratégie d'un regroupement d'individus en classes, c'est à dire constituer des classes les plus homogènes possible (les individus au sein des classes doivent se ressembler le plus possible) et à la fois des classes qui se distinguent le plus possible entre elles.

  C'est un critère qui se fonde sur la décomposition de la dispersion des individus lorsque ceux-ci sont regroupés en classes. On va préciser ce qu'il en est :

  On peut représenter sous forme de nuage de points les individus que l'on cherche à classer. Leurs positions sont déterminées par l'ensemble des caractères qui les décrivent. On calcule l'inertie totale de ce nuage qui, si les individus sont réunis en groupes, peut être décomposée en inertie à l'intérieur des classes (inertie intra) et en inertie entre les classes (inertie inter) en vérifiant la relation :

Inertie totale = Inertie inter + inertie intra

Équation 6. Décomposition de l'inertie (relation de Huygens).

Après avoir déterminé les centres de gravités des K classes (G_k) et le centre de gravité G du nuage entier, on donne les formules de calcul de ces quantités (*Équation 7*).

Inertie totale = $\sum_{i=1}^{n}(x_i - G)^2$

Dispersion entre les points et leur centre de gravité

Inertie intra = $\sum_{k=1}^{K}\sum_{i=1}^{n}(x_i - G_k)^2$

Somme par classes des dispersions entre les points et leur centre de gravité

Inertie inter = $\sum_{k=1}^{K}(G_k - G)^2$

Dispersin entre les centres de gravité de chacune des classes et le centre de gravité global.

Équation 7. Formules des composantes de l'inertie.

La *Figure 40* montre bien que si l'on agrège un individu proche du centre de gravité globale *G* à une classe, son centre de gravité s'en rapproche et l'inertie inter (entre les centres de gravités) diminue. Le critère de Ward, choisit d'agréger l'individu qui fera perdre le moins possible de cette inertie, préservant ainsi au maximum l'éloignement entre classes.

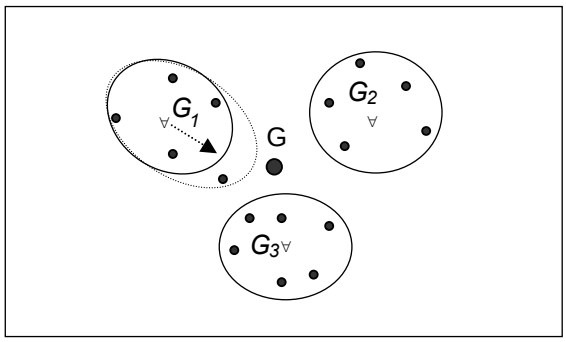


Figure 40. Agrégation d'un individu à la classe G_1 et diminution de l'inertie inter.

Cette perte d'inertie va servir d'indice de niveau de l'arbre hiérarchique permettant de décider du nombre de classes à retenir. On retiendra la partition pour laquelle l'indice de niveau de la partition suivante indique une forte perte d'inertie inter.

Le résultat d'une CAH, c'est un arbre hiérarchique qui visualise les différentes partitions successives des individus afin de choisir celle ou celles qui paraissent optimales. Nous allons par la suite décrire les résultats de la CAH des éleveurs du Breedland.

6.2.4. CAH à partir des coordonnées factorielles

Une pratique très fréquente en CAH consiste à classer les individus à partir de leurs coordonnées factorielles. On choisit de faire les regroupements des individus

à partir des axes retenus pour l'interprétation, donc aux éléments essentiels (principaux) de distinction des éleveurs en faisant abstraction de la partie difficilement interprétable ou « bruit de fond » (l'information recueillie sur les axes non significatifs). Ce procédé est très utile lorsque il existe une grande homogénéité entre les individus et qu'il est donc difficile, si l'on considère l'ensemble de l'information, d'obtenir une classification avec des groupes bien distincts.

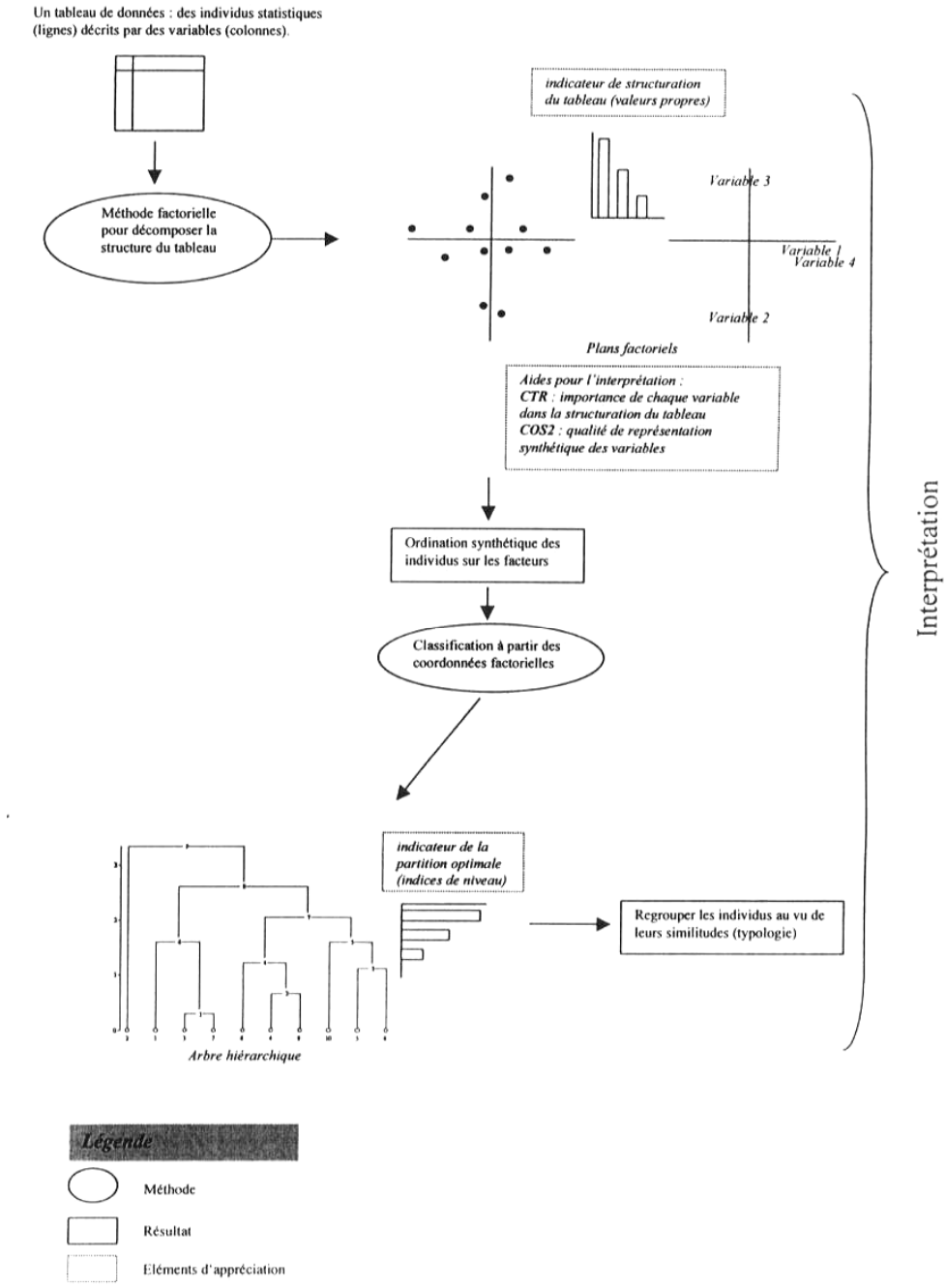


Figure 41. Synoptique d'une analyse typologique.

C'est donc souvent à partir des grands traits caractéristiques dégagés par l'analyse factorielle que les partitions sont construites. Ce qui s'inscrit bien dans la démarche générale d'analyse typologique dont l'objectif final est de construire des ensembles homogènes facilement identifiables (Figure 41).

Calcul de la matrice de distance entre éleveurs basée sur les 3 premiers axes factoriels de l'acm :

```
> res.dist <- dist(res.acm$li[,1:3], method="euclidean")
```

La CAH est réalisé avec le critère d'agrégation de ward qui favorise la formation rapide de classes « compactes » :

```
> res.hclust <- hclust(res.dist, method="ward")
> res.hclust
Call:
hclust(d = res.dist, method = "ward")

Cluster method   : ward
Distance         : euclidean
Number of objects: 50
```

6.2.5. Le diagramme des indices de niveau

```
> barplot(res.hclust$height, names.arg=c(1:49), xlab="Noeuds", ylab="indices
de niveau de regroupement", cex.names=0.7)
```

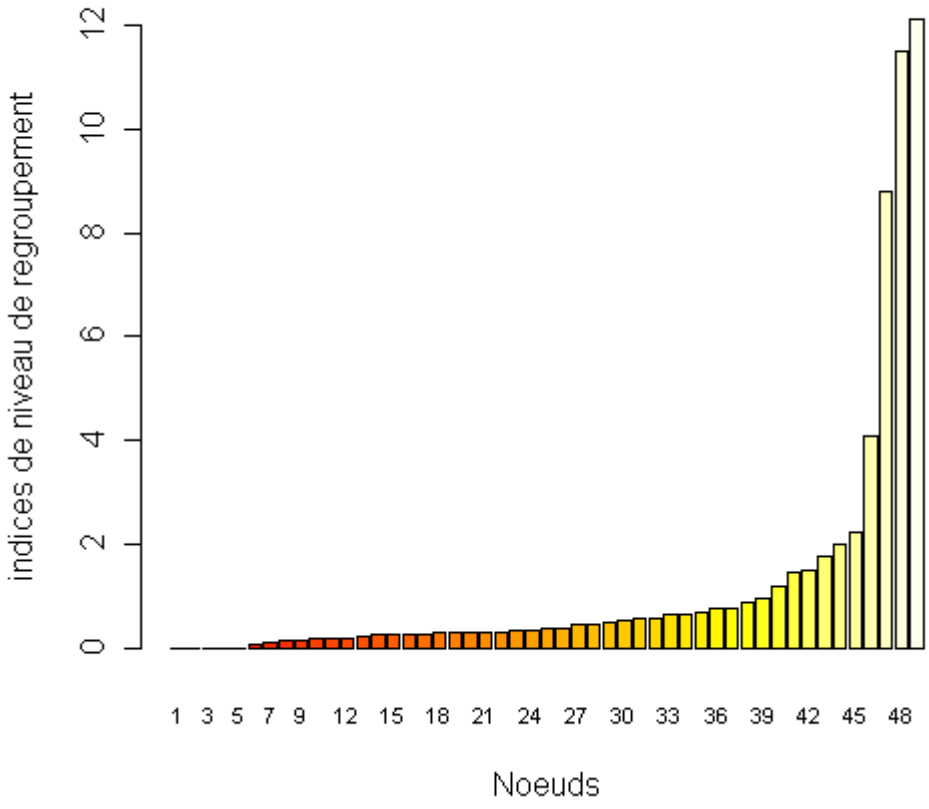


Figure 42. Diagramme des indices de niveau d'une hiérarchie.

On observe après le troisième bâton une chute importante de l'indice de niveau. Ceci signifie que l'on a regroupé deux classes d'éleveurs très éloignées. On considère que au-delà de ce bâton, les pertes sont minimes, on retient la partition en **quatre classes** (nombre de bâtons + 1). Dans le cas du choix du critère d'agrégation de Ward, le différentielle entre chaque valeur d'indice de niveau correspond à une perte d'inertie inter qui fait suite à un regroupement matérialisé par un nœud.

6.2.6. L'arbre hiérarchique ou dendrogramme

La synthèse des résultats d'une classification hiérarchique est matérialisée par l'arbre hiérarchique :

```
> plot(res.hclust, cex=0.6)
```

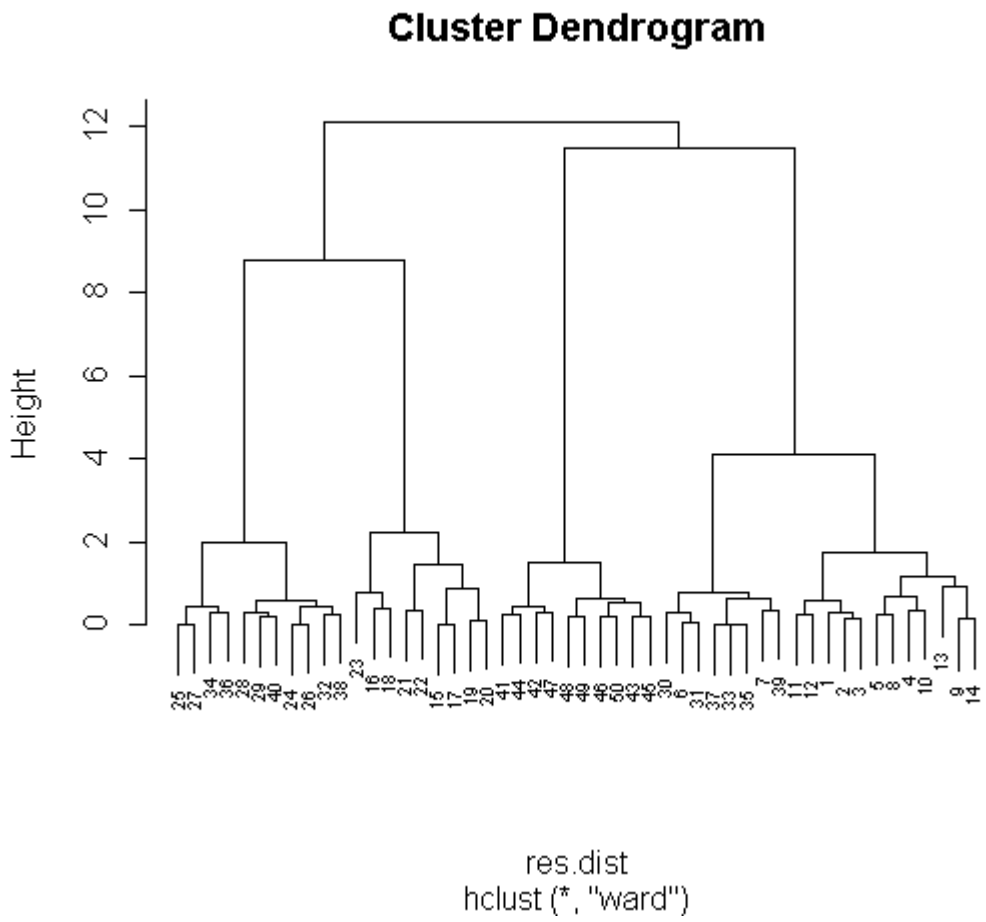


Figure 43. L'arbre hiérarchique ou dendrogramme.

Les éleveurs qui sont désignés par leur numéro en bas de l'arbre sont les feuilles et chaque regroupement d'éleveurs ou de classes d'éleveurs se symbolise par un nœud. L'axe à gauche, permet d'y lire les indices de niveau à chaque nœud. A chaque nœud, les indices représentent la perte d'inertie inter induite par les regroupements successifs et la longueur des « branches » est une bonne représentation de l'éloignement des classes. La coupure en 4 classes paraît assez évidente :

```
> partition <- cutree(res.hclust, k=4)
```

transformer la variable `partition` en facteur :

```
> partition <- as.factor(partition)
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 3 3 3 3 3 3 1 1 3 1 3 1 3 1 3
[39] 1 3 4 4 4 4 4 4 4 4 4 4
Levels: 1 2 3 4
```

6.2.7. La description des classes

Voici quelques résultats numériques et graphiques pour nous aider à qualifier les différentes classes d'élevages construites par la CAH.

Une représentation des classes sur les plans factoriels de l'ACM préliminaire autorise l'utilisation des résultats précédemment acquis pour la caractérisation de la partition :

```
> s.class(res.acm$li, partition, xax=1, yax=2, cstar=1, cellipse=1.96)
```

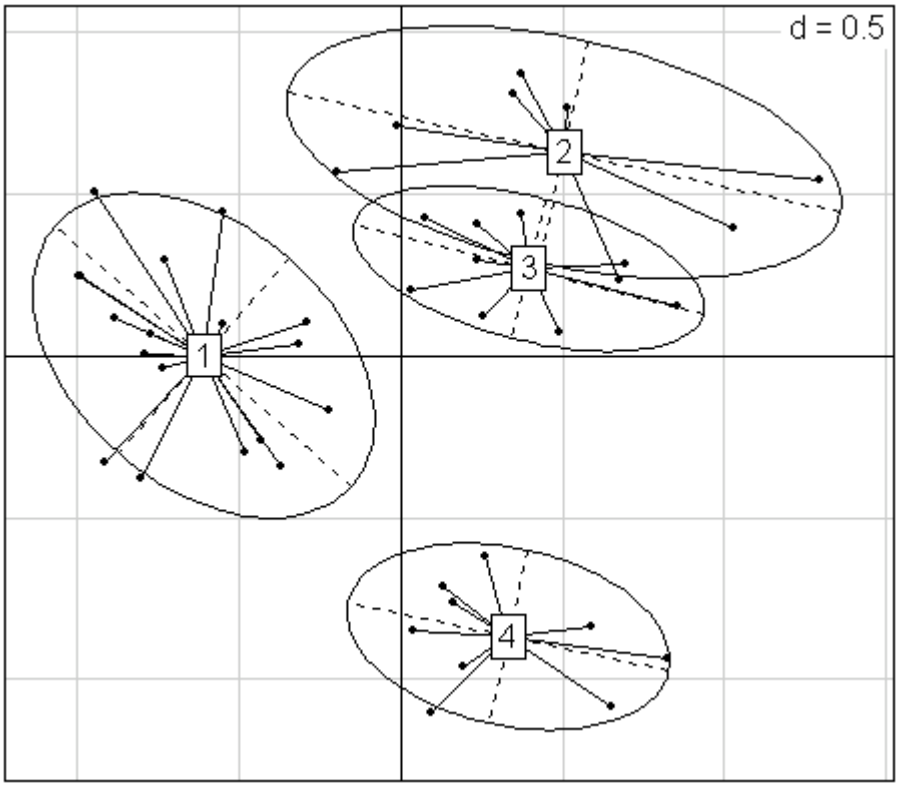


Figure 44. Représentation des classes de la partition des éleveurs sur le plan 1-2 des individus.

La partition a des spécificités régionales puisque la variable `region` discrimine très fortement les éleveurs du Breedland :

```
> table(partition, region)
         region
partition ne no sc
        1 10  4  6
        2  0  9  0
        3  0  0 11
        4  0 10  0
```

Nous disposons dans le tableau eleveurs de la variable quantitative bovinsq (nombre de bovins) :

```
> tapply(eleveurs[,"bovinsq"], partition, mean)
       1        2        3        4
87.30000 22.88889 22.36364 12.70000

> tapply(eleveurs[,"bovinsq"], partition, median)
   1    2    3    4
78.0 19.0 19.0  8.5

> tapply(eleveurs[,"bovinsq"], partition, sd)
        1         2         3         4
34.674577 15.078498  8.766672 10.382143
```

La fonction `vtest` du package `ttool` (http://forums.cirad.fr/logiciel-R/) permet d'analyser les associations entre la partition et les variables actives ou toute autre ensemble de descripteurs n'ayant pas été utilisé pour la construction des classes. Les valeurs-tests (Morineau, 1984 ; Lebart et al., 1995) mesurent pour une variable continue ou nominale sa contribution à l'originalité de la classe.

```
f <- vtest(cbind(paturage, eau, pathologies, feux, volinsecu, mouches) ~
partition, data = varsup)
f

$CALL
vtest(formula = cbind(paturage, eau, pathologies, feux, volinsecu,
    mouches) ~ partition, data = varsup)

$vtest
                        1       2       3       4
paturage.patun      0.9206  2.5981 -1.4070 -2.2685
paturage.patuo     -0.9206 -2.5981  1.4070  2.2685
eau.eaun           -0.2252 -0.6835  1.8477 -1.0133
eau.eauo            0.2252  0.6835 -1.8477  1.0133
pathologies.patho  -0.8059  3.5236 -0.4008 -1.6118
pathologies.pathon  0.8059 -3.5236  0.4008  1.6118
feux.feuxn         -1.5421  2.0606 -1.5848  2.1121
feux.feuxo          1.5421 -2.0606  1.5848 -2.1121
volinsecu.voln      0.4367  0.5563 -0.3767 -0.7643
volinsecu.volo     -0.4367 -0.5563  0.3767  0.7643
mouches.moucn      -1.5285 -0.5563  1.9037  0.7643
mouches.mouco       1.5285  0.5563 -1.9037 -0.7643

$pval
                         1      2      3      4
paturage.patun     0.3573 0.0094 0.1594 0.0233
paturage.patuo     0.3573 0.0094 0.1594 0.0233
eau.eaun           0.8218 0.4943 0.0647 0.3109
eau.eauo           0.8218 0.4943 0.0647 0.3109
pathologies.patho  0.4203 0.0004 0.6886 0.1070
pathologies.pathon 0.4203 0.0004 0.6886 0.1070
feux.feuxn         0.1230 0.0393 0.1130 0.0347
feux.feuxo         0.1230 0.0393 0.1130 0.0347
volinsecu.voln     0.6623 0.5780 0.7064 0.4447
volinsecu.volo     0.6623 0.5780 0.7064 0.4447
mouches.moucn      0.1264 0.5780 0.0570 0.4447
mouches.mouco      0.1264 0.5780 0.0570 0.4447

attr(,"class")
[1] "vtest"
```

On visualise sur un graphique les valeurs-tests des modalités rangées par classe :

```
plot(f, conf = 95, cex = 0.6)
```

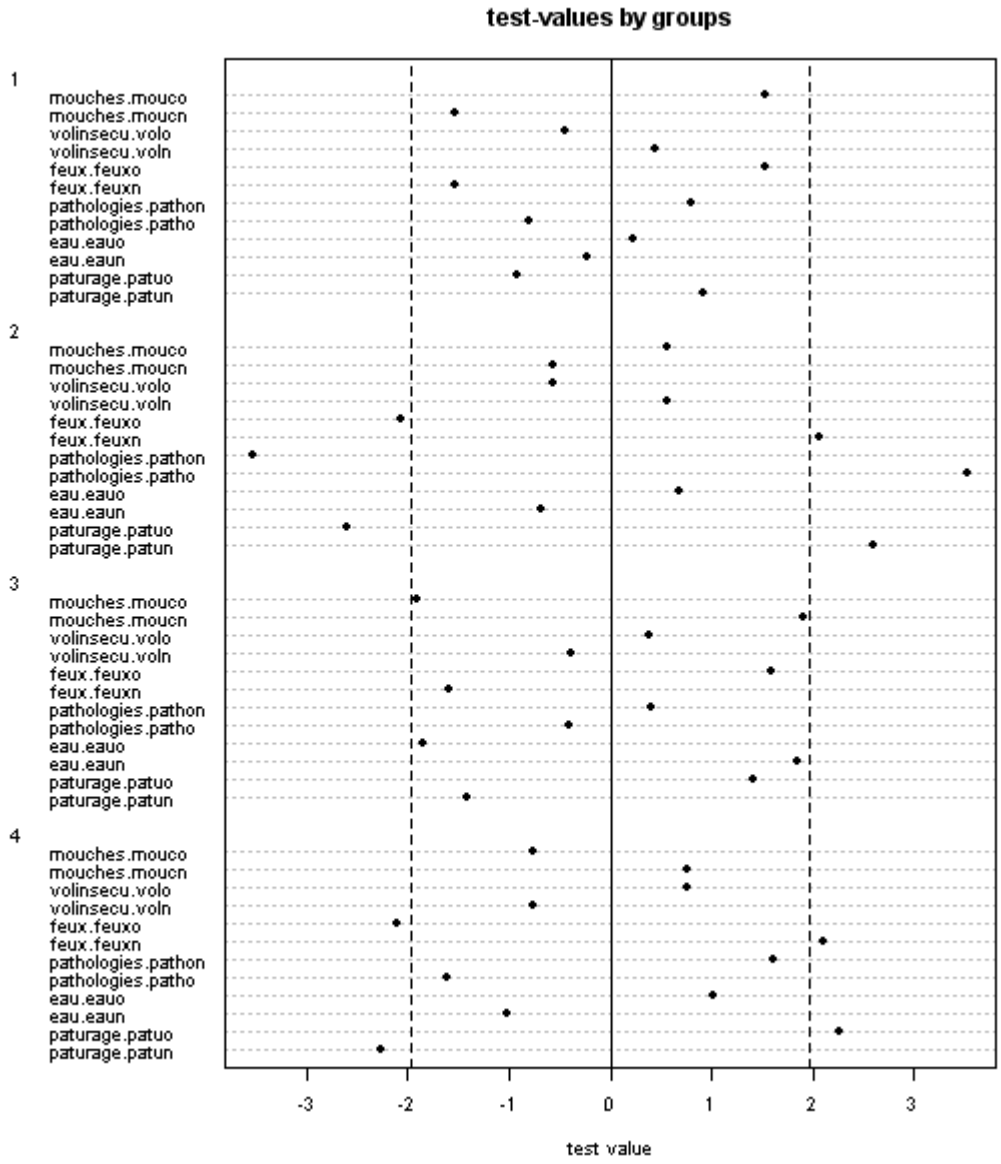


Figure 45. Description à l'aide des valeurs-tests de la la partition des éleveurs à l'aide des des réponses aux questions sur les problèmes de développement (variables supplémentaires).

Noter dans l'aide de `vtest` (`?vtest`), La mise en garde dans l'emploi des probabilités d'erreur lorsque les descripteurs ont servi à la réalisation de la classification. Il est préférable en effet dans ce cas précis de d'interpéter les classes en recourant à un simple classement des descripteurs en fonction des valeurs-tests.

6.2.8. La complémentarité analyse factorielle et classification

En analyse factorielle on veut réduire le nombre de descripteurs au profit de variables synthétiques et en classification, on veut remplacer les *n* individus par des ensembles plus faciles à décrire : les classes. Finalement les objectifs d'une analyse factorielle et d'une méthode de classification sont similaires : obtenir une idée simple des données qui transcrivent une réalité complexe.

L'AFCM et la CAH des éleveurs du Breedland est un cas typique pour illustrer la complémentarité de ces deux méthodes. On a vu comment, l'analyse factorielle nous permettait de « déblayer » le terrain pour mieux comprendre les réponses des éleveurs au questionnaire et pouvoir ainsi dégager des éléments de distinction essentiels. Puis c'est à partir de ces éléments (les facteurs) que la CAH nous a permis de classer de manière automatique les éleveurs.

On explique comment cette complémentarité peut pallier à certains inconvénients de l'analyse factorielle (Lebart et al., 1995) :

- Plans factoriels illisibles

  lorsque l'analyse factorielle est effectuée sur des milliers d'individus, les plans factoriels deviennent inextricables et illisibles. On peut simplifier, en regroupant les individus en classes que l'on projettera sur les plans factoriels.

- Robustesse

  un point s'il est très atypique peut masquer tout le reste de l'information qui aurait dû être mise en évidence par l'analyse factorielle. La classification isole ces points dans des classes sans pour autant être perturbée pour la construction des autres ensembles d'individus.

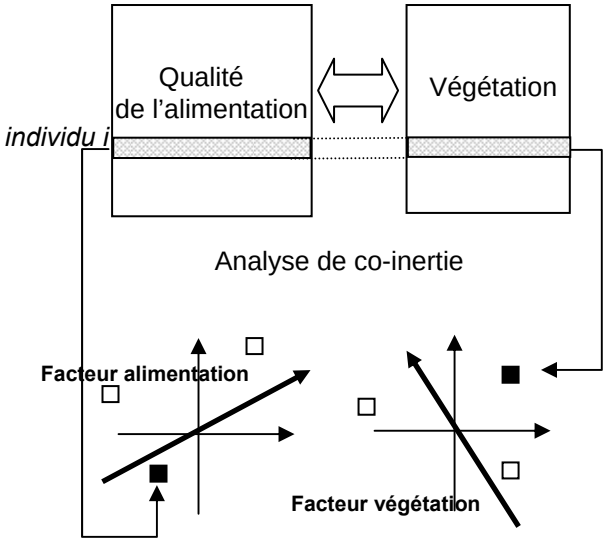
- Dimensions cachées

  les plans factoriels au-delà du plan principal (F1-F2) sont très souvent difficiles à interpréter. Pour une AFCM, les associations entre modalités y sont moins évidentes à lire car elles dépendent de celles qui ont été examinées sur les plans précédents. La classification sur facteurs prend en compte l'ensemble des dimensions principales et regroupe les individus à partir de celles-ci. On peut ainsi mettre plus facilement en avant, une classe d'individus qui se distingue par exemple sur le facteur 3 ou 4.

7. Pour aller plus loin en analyse des données

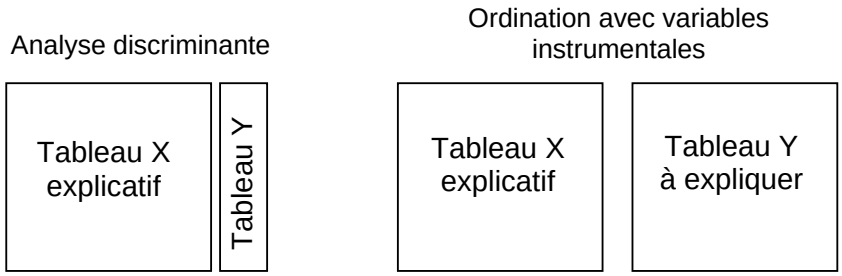
Les analyses de bases que nous avons présentées dans ce document sont un point de départ à un ensemble plus complet de méthodes, connaissant un développement très récent et qui permet de mieux tenir compte de la nature des données, de leur organisation et des objectifs d'analyse que l'on s'est fixé. La librairie ade4 met à la disposition de l'utilisateur un grand nombre de ces méthodes, à titre indicatif, on peut citer :

- Les méthodes de couplage de tableaux.



L'analyse de co-inertie permet d'analyser les relations entre 2 groupes de variables qui portent sur les mêmes individus. Par exemple, on peut vouloir identifier les liens privilégiés entre paramètres de qualité de l'alimentation et les types de végétation fréquentés par les troupeaux d'animaux sur des parcours naturels.

- Les méthodes d'analyse sur variables instrumentales



X : caractéristiques physiques ou génétiques d'un échantillon d'animaux
Y : races

X : les pratiques d'élevage
Y : pathologies survenues

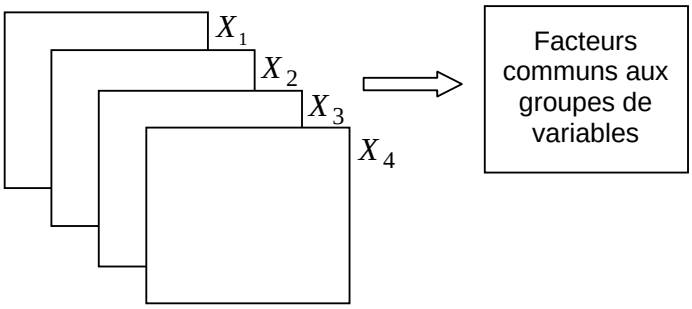
Ce groupe de méthodes confère aux tableaux des statuts différents : l'un des deux groupes de variables est par hypothèse explicatif des variations observées dans l'autre. L'analyse discriminante permet de connaître les facteurs de différenciation de deux ou plusieurs populations et l'analyse canonique des correspondances (méthode d'ordination avec variables

instrumentales) qui permet d'expliquer la dispersion des individus d'un tableau – à expliquer – à l'aide de variables contenues dans un second tableau – explicatif –.

- Les méthodes d'analyse multi-tableaux

Lorsque les mêmes individus sont décrits par des groupes de variables (un tableau par groupe de variables), l'Analyse Factorielle Multiple (AFM) permet de savoir si tous les tableaux relate la même information ou si chacun est complémentaire.

La méthode STATIS s'avère efficace, lorsque l'objectif est de définir une structure commune et ce qui différencie plusieurs tableaux qui apparient les individus ou les variables ou les deux[11]. On l'emploi notamment dans le traitement des tableaux indicés dans le temps.



Exemple d'AFM sur les tableaux suivants :

X_1 : variables définissant les structures des élevages
X_2 : variables définissant les pratiques des éleveurs
X_3 : variables définissant les pathologies survenues
X_4 : variables définissant l'environnement

[11] Un exemple d'appariement des individus et des colonnes est une succession de tableaux indicés dans le temps. On peut ainsi traiter des données d'enquête recueillies à plusieurs dates différentes.

8. Bibliographie

Escofier B., Pagès J. (1998), *Analyses factorielles simples et multiples*: *objectifs, méthodes et interprétation (3ème édition)*. Dunod, Paris.

Lebart L., Morineau A., Piron M. (1995), *Statistiques exploratoire multidimensionnelle*, 439p. Dunod, Paris.

Fenelon J.P (1981), *Qu'est-ce que l'Analyse des Données*, Lefonen, Paris.

Saporta G. (1990), *Probabilités Analyse des Données et Statistiques*, Technip, Paris.

Faye B. (1995), *Initiation à l'analyse de données*, 83p. CNPR-Ministère de l'Agriculture et de la Pêche, Lempdes.

Escofier B., Pagès J. (1997), *Initiation aux traitements statistiques. Méthodes, méthodologie*, 263p. PUR, Rennes.

Troude C., Lenoir R., Passouant M. (1993), Méthodes statistiques sous LISA – II – statistiques multivariées, 160p. CIRAD-SAR N°40/93.

Philippeau G. (1992), *Comment interpréter les résultats d'une analyse en composantes principales ?*, 61p. collection STAT-ITCF, ITCF.

Dervin C. (1992), *Comment interpréter les résultats d'une analyse factorielle des correspondances ?*, 72 p. collection STAT-ITCF, ITCF.

Jambu M. (1989), *Exploration informatique et statistique des données*, 505 p., Collection technique et scientifique des télécommunications, Dunod.

Grangé D., Lebart L. (1993*), Traitements statistiques des enquêtes*, 255p., Dunod, Paris.

Tomassone R., Danzard M., Daudin J.J, Masson J.P. (1988), Discrimination et classement, Masson, Paris.

Morineau, A. (1984). Note sur la caractérisation statistique d'une classe et les valeurs tests. Bulletin Technique du Centre de Statistique et d'Informatique Appliqués. 1, 9:12.

Outils logiciels

Les exemples présentés dans ce document s'appuient sur le logiciel d'environnement de calcul statistique et graphique **R** (version 1.6.2), sur les librairies (packages) de fonctions dédiées aux analyses de données multivariées et de représentation graphique **ade4** (version 1.00) et **mva** (version 1.6.2).

8.1. Liens internet

Nous conseillons vivement la consultation des liens web suivants qui vous permettront d'en savoir beaucoup plus sur ces outils en évolution et de télécharger l'ensemble des ressources utiles :

Le site officiel du projet de logiciel libre **R** :
http://www.r-project.org/

Le site CRAN : 'Comprehensive R Archive Network' qui compile toutes les ressources logiciels et documentaires autour de R :
http://cran.r-project.org/

Le site du projet **ade4** developpé au laboratoire d'écologie statistique de l'université Claude Bernard (Lyon 1) :
http://pbil.univ-lyon1.fr/ADE-4/ADE-4F.html

La librarie **ade-4** est téléchargeable depuis le site CRAN ou directement via le menu : 'install package from CRAN' du logiciel **R**. Ci-dessous un lien sur cette librarie pour configuration PC Windows 95 ou plus :
http://cran.r-project.org/bin/windows/contrib/ade4.zip

8.2. L'aide en ligne des fonctions des librairies ade4 et mva.

Plusieurs possibilités sont offertes pour accéder à l'aide en ligne des fonctions des librairies **ade4** et **mva** :

1. via le prompt du logiciel R :
> help(--nom de la fonction--)

2. via le menu : Help/R Language (html) . Aide en ligne html nécessitent un navigateur web installée sur le PC.

3. via le fichier d'aide : ade4.chm situé dans le répertoire d'installation de la librairie ade4 : R\rw1062\library\ade4\chtml

8.3. Liste des fonctions de la librairie ade4

```
> library(help=ade4)


                Information sur le package 'ade4'

Description :

Package:        ade4
Version:        1.4-5
Date:           2007/10/12
Title:          Analysis of Ecological Data : Exploratory and Euclidean
methods in Environmental sciences
Author:         Daniel Chessel, Anne-Beatrice Dufour <dufour@biomserv.univ-
lyon1.fr> and Stephane Dray <dray@biomserv.univ-lyon1.fr>, with
contributions from Jean R. Lobry,
                Sebastien Ollier, Sandrine Pavoine and Jean Thioulouse.
Maintainer:     Simon Penel <penel@biomserv.univ-lyon1.fr>
Suggests:       waveslim, splancs, MASS, maptools, spdep, pixmap, ape,
tripack, ade4TkGUI
Description:    Multivariate data analysis and graphical display.
License:        GPL version 2 or newer
URL:            http://pbil.univ-lyon1.fr/ADE-4, Mailing list:
http://listes.univ-lyon1.fr/wws/info/adelist
Packaged:       Fri Oct 12 13:52:03 2007; penel
Built:          R 2.6.0; i386-pc-mingw32; 2007-10-12 19:48:03; windows

Index :

PI2newick               Import data files from Phylogenetic
                        Independance Package
RV.rtest                Monte-Carlo Test on the sum of eigenvalues of
                        a co-inertia analysis (in R).
RVdist.randtest         Tests of randomization on the correlation
                        between two distance matrices (in R).
abouheif.eg             Phylogenies and quantitative traits from
                        Abouheif
acacia                  Spatial pattern analysis in plant communities
ade4toR                 Format Change Utility
aminoacyl               Codon usage
amova                   Analysis of molecular variance
apis108                 Allelic frequencies in ten honeybees
                        populations at eight microsatellites loci
ardeche                 Fauna Table with double (row and column)
                        partitioning
area.plot               Graphical Display of Areas
arrival                 Arrivals at an intensive care unit
as.taxo                 Taxonomy
atlas                   Small Ecological Dataset
atya                    Genetic variability of Cacadors
avijons                 Bird species distribution
avimedi                 Fauna Table for Constrained Ordinations
aviurba                 Ecological Tables Triplet
bacteria                Genomes of 43 Bacteria
banque                  Table of Factors
baran95                 African Estuary Fishes
between                 Between-Class Analysis
bf88                    Cubic Ecological Data
bicenter.wt             Double Weighted Centring
bordeaux                Wine Tasting
bsetal97                Ecological and Biological Traits
buech                   Buech basin
butterfly               Genetics-Ecology-Environment Triple
cailliez                Transformation to make Euclidean a distance
                        matrix
capitales               Road Distances
carni19                 Phylogeny and quantative trait of carnivora
carni70                 Phylogeny and quantitative traits of carnivora
carniherbi49            Taxonomy, phylogenies and quantitative traits
                        of carnivora and herbivora
casitas                 Enzymatic polymorphism in Mus musculus
cca                     Canonical Correspondence Analysis
chatcat                 Qualitative Weighted Variables
```

```
chats                Pair of Variables
chazeb               Charolais-Zebus
chevaine             Enzymatic polymorphism in Leuciscus cephalus
clementines          Fruit Production
cnc2003              Frequenting movie theaters in France in 2003
coinertia            Coinertia Analysis
coleo                Table of Fuzzy Biological Traits
corkdist             Tests of randomization between distances
                     applied to 'kdist' objetcs
corvus               Corvus morphology
deug                 Exam marks for some students
disc                 Rao's dissimilarity coefficient
discrimin            Linear Discriminant Analysis (descriptive
                     statistic)
discrimin.coa        Discriminant Correspondence Analysis
dist.binary          Computation of Distance Matrices for Binary
                     Data
dist.dudi            Computation of the Distance Matrix of a
                     Statistical Triplet
dist.genet           Genetic distances from gene frequencies
dist.neig            Computation of the Distance Matrix associated
                     to a Neighbouring Graph
dist.prop            Computation of Distance Matrices of Percentage
                     Data
dist.quant           Computation of Distance Matrices on
                     Quantitative Variables
divc                 Rao's diversity coefficient also called
                     quadratic entropy
divcmax              Maximal value of Rao's diversity coefficient
                     also called quadratic entropy
dotcircle            Representation of n values on a circle
doubs                Pair of Ecological Tables
dpcoa                Double principal coordinate analysis
dudi                 Duality Diagram
dudi.acm             Multiple Correspondence Analysis
dudi.coa             Correspondence Analysis
dudi.dec             Decentred Correspondence Analysis
dudi.fca             Fuzzy Correspondence Analysis
dudi.hillsmith       Ordination of Tables mixing quantitative
                     variables and factors
dudi.mix             Ordination of Tables mixing quantitative
                     variables and factors
dudi.nsc             Non symmetric correspondence analysis
dudi.pca             Principal Component Analysis
dudi.pco             Principal Coordinates Analysis
dunedata             Dune Meadow Data
ecg                  Electrocardiogram data
ecomor               Ecomorphological Convergence
elec88               Electoral Data
escopage             K-tables of wine-tasting
euro123              Triangular Data
fission              Fission pattern and heritable morphological
                     traits
foucart              K-tables Correspondence Analysis with the same
                     rows and the same columns
friday87             Faunistic K-tables
fruits               Pair of Tables
fuzzygenet           Reading a table of genetic data (diploid
                     individuals)
gearymoran           Moran's I and Geary'c randomization tests for
                     spatial and phylogenetic autocorrelation
genet                A class of data: tables of populations and
                     alleles
granulo              Granulometric Curves
gridrowcol           Complete regular grid analysis
housetasks           Contingency Table
humDNAm              human mitochondrial DNA restriction data
ichtyo               Point sampling of fish community
inertia.dudi         Statistics of inertia in a one-table analysis
irishdata            Geary's Irish Data
is.euclid            Is a Distance Matrix Euclidean ?
julliot              Seed dispersal
jv73                 K-tables Multi-Regions
kcponds              Ponds in a nature reserve
kdist                the class of objects 'kdist' (K distance
                     matrices)
kdist2ktab           Transformation of K distance matrices (object
```

	'kdist') into K Euclidean representations (object 'ktab')
kdisteuclid	a way to obtain Euclidean distance matrices
kplot	Generic Function for Multiple Graphs in a K-tables Analysis
kplot.foucart	Multiple Graphs for the Foucart's Correspondence Analysis
kplot.mcoa	Multiple Graphs for a Multiple Co-inertia Analysis
kplot.mfa	Multiple Graphs for a Multiple Factorial Analysis
kplot.pta	Multiple Graphs for a Partial Triadic Analysis
kplot.sepan	Multiple Graphs for Separated Analyses in a K-tables
kplot.statis	Multiple Graphs of a STATIS Analysis
krandtest	Class of the Permutation Tests (in C).
ktab	the class of objects 'ktab' (K-tables)
ktab.data.frame	Creation of K-tables from a data frame
ktab.list.df	Creating a K-tables from a list of data frames.
ktab.list.dudi	Creation of a K-tables from a list of duality diagrams
ktab.match2ktabs	STATIS and Co-Inertia : Analysis of a series of paired ecological tables
ktab.within	Process to go from a Within Analysis to a K-tables
lascaux	Genetic/Environment and types of variables
lingoes	Transformation of a Distance Matrix for becoming Euclidean
lizards	Phylogeny and quantitative traits of lizards
macaca	Landmarks
macon	Wine Tasting
mafragh	Phyto-Ecological Survey
mantel.randtest	Mantel test (correlation between two distance matrices (in C).)
mantel.rtest	Mantel test (correlation between two distance matrices (in R).)
maples	Phylogeny and quantitative traits of flowers
mariages	Correspondence Analysis Table
mcoa	Multiple CO-inertia Analysis
meau	Ecological Data : sites-variables, sites-species, where and when
meaudret	Ecological Data : sites-variables, sites-species, where and when
mfa	Multiple Factorial Analysis
microsatt	Genetic Relationships between cattle breeds with microsatellites
mjrochet	Phylogeny and quantitative traits of teleos fishes
mld	Multi Level Decomposition of unidimensional data
mollusc	Faunistic Communities and Sampling Experiment
monde84	Global State of the World in 1984
morphosport	Athletes' Morphology
mstree	Minimal Spanning Tree
multispati	Multivariate spatial analysis
multispati.randtest	Multivariate spatial autocorrelation test (in C)
multispati.rtest	Multivariate spatial autocorrelation test
neig	Neighbourhood Graphs
newick.eg	Phylogenetic trees in Newick format
newick2phylog	Create phylogeny
niche	Method to Analyse a pair of tables : Environmental and Faunistic Data
njplot	Phylogeny and trait of bacteria
olympic	Olympic Decathlon
oribatid	Oribatid mite
orthobasis	Orthonormal basis for orthonormal transform
orthogram	Orthonormal decomposition of variance
ours	A table of Qualitative Variables
palm	Phylogenetic and quantitative traits of amazonian palm trees
pap	Taxonomy and quantitative traits of carnivora
pcaiv	Principal component analysis with respect to instrumental variables
pcaivortho	Principal Component Analysis with respect to orthogonal instrumental variables

pcoscaled	*Simplified Analysis in Principal Coordinates*
perthi02	*Contingency Table with a partition in Molecular Biology*
phylog	*Phylogeny*
plot.phylog	*Plot phylogenies*
presid2002	*Results of the French presidential elections of 2002*
procella	*Phylogeny and quantitative traits of birds*
procuste	*Simple Procruste Rotation between two sets of points*
procuste.randtest	*Monte-Carlo Test on the sum of the singular values of a procustean rotation (in C).*
procuste.rtest	*Monte-Carlo Test on the sum of the singular values of a procustean rotation (in R).*
pta	*Partial Triadic Analysis of a K-tables*
quasieuclid	*Transformation of a distance matrice to a Euclidean one*
randtest	*Class of the Permutation Tests (in C).*
randtest.amova	*Permutation tests on an analysis of molecular variance (in C).*
randtest.between	*Monte-Carlo Test on the between-groups inertia percentage (in C).*
randtest.coinertia	*Monte-Carlo test on a coinertia analysis (in C).*
randtest.discrimin	*Monte-Carlo Test on a Discriminant Analysis (in C).*
rankrock	*Ordination Table*
reconst	*Reconstitution of Data from a Duality Diagram*
rhone	*Physico-Chemistry Data*
rlq	*RLQ analysis*
rpjdl	*Avifauna and Vegetation*
rtest	*Class of the Permutation Tests (in R).*
rtest.between	*Monte-Carlo Test on the between-groups inertia percentage (in R).*
rtest.discrimin	*Monte-Carlo Test on a Discriminant Analysis (in R).*
s.arrow	*Plot of the factorial maps for the projection of a vector basis*
s.chull	*Plot of the factorial maps with polygons of contour by level of a factor*
s.class	*Plot of factorial maps with representation of point classes*
s.corcircle	*Plot of the factorial maps of a correlation circle*
s.distri	*Plot of a frequency distribution*
s.hist	*Display of a scatterplot and its two marginal histograms*
s.image	*Graph of a variable using image and contour*
s.kde2d	*Scatter Plot with Kernel Density Estimate*
s.label	*Scatter Plot*
s.match	*Plot of Paired Coordinates*
s.traject	*Trajectory Plot*
s.value	*Representation of a value in a graph*
santacatalina	*Indirect Ordination*
sarcelles	*Array of Recapture of Rings*
scalewt	*Centring and Scaling a Matrix of Any Weighting*
scatter	*Scatter Plot*
scatter.acm	*Plot of the factorial maps in multiple correspondence analysis*
scatter.coa	*Plot of the factorial maps for a correspondence analysis*
scatter.dudi	*Plot of the Factorial Maps*
scatter.fca	*Plot of the factorial maps for a fuzzy correspondence analysis*
sco.boxplot	*Representation of the link between a variable and a set of qualitative variables*
sco.distri	*Representation by mean- standard deviation of a set of weight distributions on a numeric score*
sco.quant	*Graph to Analyse the Relation between a Score and Quantitative Variables*
score	*Graphs for One Dimension*
score.acm	*Graphs to study one factor in a Multiple Correspondence Analysis*
score.coa	*Graphs to analyse a factor in a correspondence analysis*
score.mix	*Graphs to Analyse a factor in a Mixed Analysis*

score.pca	*Graphs to Analyse a factor in PCA*
seconde	*Students and Subjects*
sepan	*Separated Analyses in a K-tables*
skulls	*Morphometric Evolution*
statis	*STATIS, a method for analysing K-tables*
steppe	*Transect in the Vegetation*
supcol	*Projections of Supplementary Columns*
suprow	*Projections of Supplementary Rows*
symbols.phylog	*Representation of a quantitative variable in front of a phylogenetic tree*
syndicats	*Two Questions asked on a Sample of 1000 Respondents*
t3012	*Average temperatures of 30 French cities*
table.cont	*Plot of Contingency Tables*
table.dist	*Graph Display for Distance Matrices*
table.paint	*Plot of the arrays by grey levels*
table.phylog	*Plot arrays in front of a phylogenetic tree*
table.value	*Plot of the Arrays*
tarentaise	*Mountain Avifauna*
taxo.eg	*Examples of taxonomy*
tintoodiel	*Tinto and Odiel estuary geochemistry*
tithonia	*Phylogeny and quantitative traits of flowers*
tortues	*Morphological Study of the Painted Turtle*
toxicity	*Homogeneous Table*
triangle.class	*Triangular Representation and Groups of points*
triangle.plot	*Triangular Plotting*
trichometeo	*Pair of Ecological Data*
ungulates	*Phylogeny and quantitative traits of ungulates.*
uniquewt.df	*Elimination of Duplicated Rows in a Array*
variance.phylog	*The phylogenetic ANOVA*
veuvage	*Example for Centring in PCA*
westafrica	*Freshwater fish zoogeography in west Africa*
within	*Within Analyses*
within.pca	*Normed within Principal Component Analysis*
witwit.coa	*Internal Correspondence Analysis*
worksurv	*French Worker Survey (1970)*
yanomama	*Distance Matrices*
zealand	*Road distances in New-Zealand*

Fiche d'enquête - Les éleveurs du Breedland

N° de fiche : Date :
Nom de l'enquêteur :
Région (1-Nord-Est, 2-Nord-Ouest, 3-Sud) :

1 - Identification de l'éleveur-propriétaire

Nom de l'éleveur : Ethnie :
Commune : (1-Wala, 2-Wolo, 3-Wulu)
(1-Farmtown, 2-Breedtown, 3-Reartown)

Combien avez-vous de femmes ? ____
Combien avez-vous d'enfants ? ____

Pratiques de déplacement : sédentaire transhumant transhumant partiellement
(une seule réponse possible)

Avez-vous une autre activité ? oui non
Si oui,
est-elle liée à : l'agriculture au commerce à l'artisanat (une seule réponse possible)

2 - Taille et diversité du troupeau

Nombre total de bovins : ____
(si l'éleveur possède plusieurs troupeaux, comptabiliser tous les animaux)

Les bovins sont élevés pour : la viande le lait les deux

Se reporter page 2 pour la saisie des informations sur chaque animal

Possédez-vous des petits ruminants : oui non
Si oui, combien : ____

3 - Pratiques d'élevage

Qui garde le troupeau ? une personne de la famille
une personne salariée

Mise en lots des bovins : oui non

Vaccinez-vous vos bêtes : systématiquement non parfois

Fréquence de vermifugation de vos bêtes : ____

3 - Les problèmes de développement

Quels sont les problèmes que vous rencontrez au quotidien ?
(plusieurs réponses possibles)
A ordonner du plus contraignant au moins contraignant (4 niveaux)

	1	2	3	4
manque de pâturages				
rareté de l'eau				
Pathologies animales				
gêne des feux de brousse				
vols/insécurité				
mouches				

Le troupeau : informations individuelles

N°	Sexe (M/F)	Age	Race (N'Dama, Baoule, M'Boro)	NEC[12] (1-5)	Nb mises bas (99 si Sexe=M)	Origine (1-Don, 2-Prêt, 3-Achat, 4-Renouvellement)	Destinée[13]

[12] NEC : note d'état corporel
[13] Destinée : 1-Reproduction, 2-Vente, 3-Autoconso., 4-Vente ou Autoconso., 5-Non déterminée.